Assessment of Malingering With Repeat Forensic Evaluations: Patient Variability and Possible Misclassification on the SIRS and Other Feigning Measures

Richard Rogers, PhD, Michael J. Vitacco, PhD, and Samantha J. Kurus, BA

Patients with Axis I disorders often fluctuate markedly in their clinical presentations in forensic and other professional settings. Although such fluctuations could suggest ineffectual efforts at malingering, more likely explanations include confusion or poor insight into psychopathology, imprecision in the assessment methods, or actual changes in symptomatology. An important concern is whether such fluctuations, common in repeat forensic evaluations, might lead to false-positive results—specifically, the misclassifications of patients as malingerers. We used the Structured Interview of Reported Symptoms (SIRS) to examine the effects of repeat administration of the interview on 52 likely genuine forensic inpatients. As expected, test-retest correlations for individual SIRS scales were highly variable. Despite this variability, the magnitude of differences remained small. The SIRS produced no errors in the classification of likely genuine forensic inpatients at the first or the repeat administrations. Implications of variable clinical presentations for other feigning measures are considered.

J Am Acad Psychiatry Law 38:109–14, 2010

According to the adaptational model,[1] malingering is conceptualized as a specific response style to an adverse set of circumstances. As a situational response style, malingering is not viewed as a stable trait or enduring characteristic of feigning individuals. Therefore, the notion "once a malingerer, always a malingerer" is now considered a basic myth of malingering (Ref. 2, p 7). As in the insanity defense, for example, a substantial minority of criminal defendants (see Rogers and Shuman[3]) are motivated to malinger in an effort to be found not guilty by reason of insanity (NGRI). If successfully acquitted as NGRI, the motivation to malinger ceases immediately. Those found NGRI, whether malingering or not, are likely to adopt an opposite response set of simulated adjustments in an effort to secure their release from a secure forensic hospital. As illustrated by the NGRI example, most cases of malingering are situationally determined and goal-specific.

This brief analysis addresses a critical concern of repeat forensic evaluations. Does the marked variability in clinical presentation, common among patients with genuine psychotic and other Axis I diagnoses, lead to false-positive findings (misclassifying genuine patients as malingerers) on repeat administration of feigning measures? Such grave errors could lead to unwarranted conclusions about malingering and undermine the foundation of a forensic report.

For forensic assessments, reliable and reproducible measurements are the *sine qua non* of standardized measures. With a primary focus on legitimacy of the examinee's current clinical presentation, feigning measures typically emphasize interrater reliability and the reliability of individual scores (standard error of measurement,[4] or SEM). As a brief review, interrater reliability assesses the level of agreement between independent evaluators; high correlations pro-

Dr. Rogers is Professor of Psychology, University of North Texas, Denton, TX; Dr. Vitacco is Associate Director of Research and Ms. Kurus is Research Assistant, Mendota Mental Health Institute, Madison, WI. Address correspondence to: Richard Rogers, PhD, University of North Texas, Department of Psychology, 1155 Union Circle #311280, Denton, TX 76203-5017. E-mail: rogers_1@hotmail.com.

Disclosures of financial or other potential conflicts of interest: Dr. Rogers receives royalties for the SIRS.

Table 1 Comparison of SIRS Scores Across Time With NGRI Inpatients

	Time 1		Time 2		Differences			Reliability	
Scale	*M*	*SD*	*M*	*SD*	$\lvert$M Diff.$\rvert$	*F*	*d*	SEM-1	SEM-2
RS*	0.75	1.28	0.56	0.96	0.19	0.75	0.17	0.18	0.13
SC*	0.77	1.44	0.77	1.50	0.00	0.00	0.00	0.20	0.21
IA*	0.13	0.49	0.31	0.78	−0.18	1.84	0.27	0.07	0.11
BL†	0.94	1.64	0.62	1.39	0.32	1.21	0.21	0.23	0.19
SU†	4.46	4.79	4.15	5.48	0.31	0.09	0.06	0.66	0.76
SEL†	4.54	4.29	3.85	4.54	0.69	0.64	0.16	0.60	0.63
SEV†	0.87	1.90	0.92	2.09	−0.05	0.02	0.03	0.26	0.29
RO*	0.17	0.51	0.04	0.28	−0.13	2.77	0.32	0.07	0.38
$\mathrm{M}_{\mathrm{unlikely}}$	0.46	0.93	0.42	0.88	0.04	1.34	0.19	0.13	0.21
$\mathrm{M}_{\mathrm{amplified}}$	2.70	3.16	2.39	3.38	0.31	0.49	0.12	0.44	0.47

|M diff.|, the absolute mean difference; SEM-1, SEM for Time 1; SEM-2, SEM for Time 2; RS, rare symptoms; SC, symptom combinations; IA, improbable and absurd symptoms; BL, blatant symptoms; SU, subtle symptoms; SEL, selectivity of symptoms; SEV, severity of symptoms; RO, reported vs. observed symptoms. All *F* ratios are nonsignificant.
*Scale uses an unlikely detection strategy.
†Scale uses an amplified detection strategy.

vide strong evidence of good interrater reliability. SEMs estimate the variability of individual scores; therefore, low SEMs provide strong evidence regarding the reliability of individual scores.

For the study of repeat forensic examinations, test-retest reliability (i.e., reproducibility of scores over time) must be considered for all psychometric methods. In genuine patients, test-retest reliability will be adversely affected by fluctuations in clinical presentation. For malingerers, test-retest reliability has only marginal relevance; some feigners do not recall their fabricated symptoms, whereas others make no effort to be consistent, believing that variability is a hallmark of severe psychological impairment.

The MMPI-2[5] and PAI[6] are multiscale inventories with feigning scales that are extensively researched and commonly used in the assessment of malingering. In focusing on the test-retest reliability of their feigning scales, however, their professional manuals[7,8] report data only on presumably unimpaired community samples rather than clinical or forensic populations. Even so, the reliability coefficients tend to be moderate, with many in the 0.7 range. When examined in clinical populations over long intervals, estimates of test-retest reliability become much more modest.[9]

Besides standardized measures, such as the MMPI-2 and PAI, the assessment of malingering has been facilitated by the development of specialized measures that are specifically designed to assess feigned mental disorders by different detection strategies. Among these measures, the Structured Interview of Reported Symptoms, or SIRS,[10] has been widely accepted by forensic experts as a well-validated measure,[11] commonly used in forensic practice.[12] The SIRS has eight primary scales (Table 1) that are used to assess feigned mental disorders and that employ two general categories: unlikely (i.e., typically bogus symptoms almost never reported by genuine patients) or amplified (i.e., typically legitimate symptoms reported at relatively low levels by genuine patients and much higher levels by feigners) strategies.[13]

For use in forensic practice, the SIRS has outstanding interrater reliabilities for its primary scales, ranging from 0.93 to 1.00 in the original validation,[10] and from 0.95 to 1.00 in more recent research.[14] In addition, the SEMs of primary scales are very low (mean (M) = 0.51) indicating high reliabilities of individual scores.[15] The current study addressed whether clinical variability in genuine forensic inpatients affects the usefulness of the SIRS in repeat forensic evaluations. Because feigners are not expected to be consistent, we focus on test-retest reliability of the SIRS in genuine patients from an inpatient forensic sample. To minimize the likelihood of malingering, we selected hospitalized patients adjudicated NGRI based on the premise that most of these participants would be seeking release from a maximum-security facility because of their improved clinical status. However, we recognized that a small number might be feigning for specific goals (e.g., desired medications). Therefore, we screened these inpatients with the Miller Forensic Assessment of Symptoms Test (M-FAST)[16] to remove potential malingerers.

Method

Participants

The initial sample was composed of 55 NGRI inpatients residing in a secure forensic unit at Mendota Mental Health Institute (Madison, WI). As will be described, three patients were excluded from the study because they exceeded the M-FAST cutoff score for possible malingering. Therefore, the final sample consisted of 52 inpatients ranging in age from 21 to 64 (M = 38.10, standard deviation (SD) = 10.10). Ethnically, the final sample was moderately diverse with 36 (69.2%) European Americans, 12 (23.1%) African Americans, 3 (5.8%) Hispanic Americans, and 1 (1.9%) Native American. The majority of patients ($n = 43$; 82.69%) had a diagnosed psychotic disorder; however, other diagnoses included mood and personality disorders, and all patients warranted multiple diagnoses.

Measures

The M-FAST[16] is a 25-item structured interview designed to screen for feigned mental disorders. These items are summed for a total score; the cutoff score ≥6 is "highly suggestive of malingered psychopathology" (Ref. 16, p 12). The M-FAST has demonstrated excellent psychometric properties in both its reliability and validity in forensic populations.[17]

The SIRS[10] is a 172-item structured interview designed to assess for feigned and related response styles. As described earlier, the SIRS has been validated and is a commonly used criterion measure for known-group studies of malingering.[14] Excellent interrater reliability has been validated at this facility.[18] Its primary scales have excellent discriminant validity, with large effect sizes, between feigning and genuine patient samples. To minimize false positives, scale scores are categorized as genuine, indeterminate, probable feigning, and definite feigning. Determinations of feigned mental disorders are based on one or more scales in the definite feigning range, or three or more scales in the probable feigning range. In marginal cases (one or two scales in the probable feigning range), a total SIRS score >76 can be applied.

Procedure

The study was approved by the Institutional Review Board at Mendota Mental Health Institute. Patients were recruited via sign-up sheets on their individual units. The general parameters of the study were explained to those expressing interest in participating. Signed informed consent was obtained from patients before their enrollment. As part of their consent, they were informed that they could discontinue the study at any time without negative consequences.

The measures were administered under instructions to be honest and forthcoming. The specific instructions included the following: "Please tell us about your current symptoms and psychological problems. Don't make them any better or worse than they are. Some forensic patients hide psychological problems or 'play up' their symptoms to get more help and attention. We need you to tell us just the way it is."

The study was completed in two distinct phases. In Phase 1, the patients were given instructions to be honest and forthcoming, and the M-FAST was then administered as a screen for possible malingering. As noted, three patients scoring 6 or higher on the M-FAST were eliminated from subsequent participation. The SIRS was administered to the remaining 52 patients under the same instructions.

Phase 2 involved the second administration of the SIRS, with the same instructions as before. The interval between interviews was approximately 10 days (M = 11.6, SD = 1.54). Immediately after completion of the study, the participants were asked follow-up questions. As a manipulation check, they were asked to repeat their instructions. Most (n = 47) recalled the instructions accurately; when prompted, however, all reported that they had presented themselves honestly. Given their reported involvement and success, none were removed from the subsequent analyses.

Results

As predicted, SIRS scales using unlikely detection strategies had very low rates of reporting that were close to zero (see Table 1) for both Times 1 (M = 0.46) and 2 (0.42), resulting in a minuscule mean difference of 0.04. Scales using amplified detection strategies were still low (M = 2.70 and 2.39, respectively). As expected, the F ratios were nonsignificant, and effect sizes were minimal (i.e., M Cohen's $d <$ 0.20; Table 1).

An important finding was the very low SEMs across Times 1 and 2, which indicated a high reliability for individual scores. As summarized in Table 1, SIRS scales with unlikely (M = 0.13 and 0.21) were even smaller than amplified (M = 0.44 and 0.47)

Table 2 SIRS Primary Scales: Reliability and Classificatioin

	Test-Retest Reliability and Consistency				
	Scales		Classification		
	Correlation	% ± |2|	Concordance (%)	κ	Yule's *Q*
RS	0.36*	92.3	98.1	0.66†	1.00
SC	0.70†	96.2	100.0	1.00†	1.00
IA	0.46†	98.1	100.0	1.00†	1.00
BL	0.54†	88.5	100.0	1.00†	1.00
SU	0.90†	75.0	94.2	0.38†	1.00
SEL	0.88†	78.6	98.1	0.66†	1.00
SEV	0.84†	92.3	98.1	0.66†	1.00
RO	−0.04	98.1	100.0	1.00†	1.00
$M_{unlikely}$	0.38	96.2	99.5	.92	1.00
$M_{amplified}$	0.79	83.6	97.6	.68	1.00

Classifications are nonfeigning (honest and indeterminate) and feigning (probable and definite); % ± |2| is the percentage for which absolute difference in repeated administrations is 2 points or less.
$^*p \leq 0.05$
$^\dagger p \leq 0.001$.

detection strategies. Using 95 percent confidence limits, practitioners can be assured that most of their recorded scores varied from 1 to 1.5 points from the actual or "true" score. This reliability of individual scores is exceptional.

Correlations for test-retest reliability when unlikely detection strategies were used were constrained by the floor effect (i.e., very low scores) with the modal score for both Times 1 and 2 being 0. As a result, the correlations fell in the low to moderate range (Table 2). The one exception was RO (Reported vs. Observed symptoms), which evidenced a negligible correlation of −0.04. However, the RO scale incorporates the practitioner's current observations of the examinee's behavior during the SIRS administration. This negligible correlation is understandable, because current behavioral observations are likely to fluctuate between Phases 1 and 2, and this scale with genuine patients often experiences a floor effect, with most scores at 0 at Times 1 (86.5%) and 2 (98.1%). In cases of restricted range, the consistency of scores across time is a useful indicator regarding the reproducibility of SIRS scales. Given that the highest score on all SIRS primary scales is 2, we used the absolute difference of ±2 (i.e., |2|) as an important benchmark. Averaging across the SIRS scales using unlikely detection strategies, nearly all scales (96.2%) exhibited very high consistency across SIRS administrations.

SIRS scales using amplified detection strategies had moderate (0.54) to high (0.84, 0.88, and 0.90) correlations between Times 1 and 2. As summarized in Table 2, the consistency rates (i.e., ± 2) were also high for the two scales (BL, Blatant Symptoms; and SEV, Severity of Symptoms) with low average scores ($M < 1.00$). For the remaining two scales (SU, Subtle Symptoms; and SEL, Selectivity of Symptoms) with somewhat higher average scores ($M > 4.00$), the consistency rates were still substantial at 75.0 percent and 78.6 percent, respectively.

The crucial test of SIRS stability is whether the scales are consistent in their classification of nonfeigning and feigning (Table 2). The overall concordance rates were exceptionally high for both unlikely (99.5%) and amplified (97.6%) strategies. We also used two coefficients of agreement: the κ statistic and Yule's *Q*. The κ statistic, with estimates that are suppressed by very low base rates, produced generally high estimates for unlikely SIRS scales but more variability for the amplified SIRS scales. Because Yule's *Q* makes no assumptions about probabilities, it provides a more interpretable measure of agreement with very low base rates. It produced consistently high coefficients across all SIRS scales. Finally, we examined the SIRS classification rules based on these primary scales and the total SIRS score. All SIRS protocols were correctly classified as genuine on both administrations, resulting in a concordance of 100 percent ($\kappa = 1.00$; Yule's $Q = 1.00$).

Discussion and Conclusions

Because of severe Axis I disorders, patients are not expected to be stable in their clinical presentations.

Repeat forensic referrals of Workers' Compensation cases revealed very little congruence in MMPI-2 code types across an interval of 21.3 months.[19] Despite presumably chronic conditions warranting continued disability coverage, more than half (62.3%) of well-defined code types were different with repeat forensic evaluations. Even with short intervals and nonforensic referrals, discrepancies commonly occur. Harrell and his colleagues[20] found that one-third of well-defined code types changed with repeat assessment after a short interval averaging only 7.9 days. The interesting question is whether this marked variability in clinical presentations (e.g. different code types) will also be observed on feigning indicators. Although initial data with the PAI suggested little variability in feigning indicators across time, a methodological flaw (i.e., exclusion of all patients with significantly elevated PAI feigning indicators) contributed to this finding.[21]

The current findings are the result of the first rigorous investigation of the stability of feigning indicators among forensic inpatients with genuine disorders. The results indicate that the SIRS primary scales remain stable with accurate and consistent scores (see SEM-1 and SEM-2 in Table 1). Although correlations for unlikely detection strategies were affected by the severely restricted range, the absolute difference between test administrations remained small. Looking beyond scores, the SIRS concordance rates (M of 97.6% and 99.5%) for feigning versus nonfeigning were exceptionally high. When the SIRS classification for feigned mental disorders was applied, 100 percent of the NGRI inpatient sample was consistently classified as nonfeigning across repeat administrations. Forensic practitioners can have a high level of confidence in the stability of SIRS scores for repeat assessments.

Practitioners must consider whether other measures used to evaluate feigned mental disorders will show similar stability for repeat forensic evaluations. In the absence of test-retest data using clinical populations for the MMPI-2 and PAI, we offer the following three recommendations:

> Low standard errors of measurement (SEMs) and 95 percent confidence limits provide good evidence regarding the accuracy and reproducibility of individual scores, although they do not address changes in individual patients across time. However, large ranges for 95 percent confidence limits indicates imprecise measures and suggest that low stability estimates will be observed.[22] As a practical matter, forensic psychiatrists can request that psychological consults routinely include SEMs and 95 percent confidence limits for each feigning indicator.
>
> When test-retest data are provided, practitioners must carefully review the data for its clinical relevance. For example, the MMPI-2 test manual[7] provides these data on community adults. Clearly, results from presumably unimpaired participants have limited generalizability to clinical populations.[23]
>
> In evaluating test stability, the reporting of differences, averaged over scales, is plainly insufficient. The variability between administrations can be obscured by sample averages (i.e., part of the sample scoring higher on first administration being cancelled out by part of the sample scoring higher on the second administration). As a concrete example, Harrell *et al.*[20] had very small mean *t* score sample differences on clinical scales ($M = 1.20T$) across MMPI-2 administrations, yet they produced a majority of discrepant code types.

In closing, patients with Axis I disorders often produce discrepant results when tests are readministered in repeat forensic evaluations. These discrepancies cannot be taken as evidence of malingering, because they occur frequently in patients with genuine disorders. This brief paper reports the first systematic examination of whether fluctuations in clinical presentation can lead to subsequent misclassifications of feigned mental disorders. In a reassuring finding, SIRS data from a forensic inpatient sample indicated highly stable classifications across repeat administrations. Questions remain about the stability of other malingering scales, especially those with large 95 percent confidence limits.

References

1. Rogers R, Salekin R, Sewell K, *et al*: A comparison of forensic and nonforensic malingerers: a prototypical analysis of explanatory models. Law Hum Behav 22:353–67, 1998
2. Rogers R: An introduction to response styles, in Clinical Assessment of Malingering and Deception (ed 3). Edited by Rogers R. New York: Guilford Press, 2008, pp 3–13
3. Rogers R, Shuman DW: Conducting Insanity Evaluations (ed 2). New York: Guilford Press, 2001
4. Anastasi A: Psychological Testing (ed 6). New York: Macmillan, 1988

5. Rogers R, Sewell KW, Martin MA, *et al*: Detection of feigned mental disorders: a meta-analysis of the MMPI-2 and malingering. Assessment 10:160–77, 2003
6. Hawes S, Boccaccini M: Detection of overreporting of psychopathology on the Personality Assessment Inventory: a meta-analytic review. Psychol Assess 21:112–24, 2009
7. Butcher JN, Dahlstrom WG, Graham JR, *et al*: MMPI-2: Manual for Administration and Scoring. Minneapolis: University of Minnesota Press, 1989
8. Morey LC: Personality Assessment Inventory Professional Manual (ed 2). Odessa, FL: Psychological Assessment Resources, 1991
9. Munley P: Comparability of MMPI-2 scales and profiles over time. J Pers Assess 78:145–60, 2002
10. Rogers R, Bagby RM, Dickens SE: Structured Interview of Reported Symptoms (SIRS) and Professional Manual. Odessa, FL: Psychological Assessment Resources, 1992
11. Lally S: What tests are acceptable for use in forensic evaluations?—a survey of experts. Prof Psychol 34:491–8, 2003
12. Archer RP, Buffington-Vollum JK, Stredny RV, *et al*: A survey of psychological test use patterns among forensic psychologists. J Pers Assess 87:84–94, 2006
13. Rogers R: Detection strategies for malingering and defensiveness, in Clinical Assessment of Malingering and Deception (ed 3). Edited by Rogers R. New York: Guilford Press, 2008, pp 14–35
14. Rogers R: Structured interviews and dissimulation, in Clinical Assessment of Malingering and Deception (ed 3). Edited by Rogers R. New York: Guilford Press, 2008, pp 301–22
15. Rogers R, Sewell KW, Gillard N: Structured Interview of Reported Symptoms-2 (SIRS-2) and Professional Manual. Odessa, FL: Psychological Assessment Resources, 2010
16. Miller HA: M-FAST: Miller-Forensic Assessment of Symptoms Test Professional Manual. Odessa, FL: Psychological Assessment Resources, 2001
17. Smith G: Brief screening measures for the detection of feigned psychopathology, in Clinical Assessment of Malingering and Deception (ed 3). Edited by Rogers R. New York: Guilford Press, 2008, pp 323–39
18. Vitacco MJ, Rogers R, Gabel J, *et al*: An evaluation of malingering screens with competency to stand trial patients: a known-groups comparison. Law Hum Behav 31:249–60, 2007
19. Livingston R, Jennings E, Colotla V, *et al*: MMPI-2: code-type congruence of injured workers. Psychol Assess 18:126–30, 2006
20. Harrell T, Honaker L, Parnell T: Equivalence of the MMPI-2 with the MMPI in psychiatric patients. Psychol Assess 4:460–5, 1992
21. Baity M, Siefert C, Chambers A, *et al*: Deceptiveness on the PAI: a study of naive faking with psychiatric inpatients. J Pers Assess 88:16–24, 2007
22. Munley P, Germain J, Tovar-Murray D, *et al*: MMPI-2 profile code types and measurement error. J Pers Assess 82:179–88, 204
23. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education: Standards for educational and psychological testing. Washington, DC: American Educational Research Association, 1999