A Clinical Comparison, Simulation Study Testing the Validity of SIMS and IOP-29 with an Italian Sample

Abstract

The Inventory of Problems – 29 (IOP-29) was recently introduced as a brief, easy-to-use measure of non-credible mental and cognitive symptoms that may be applied to a wide variety of contexts or clinical conditions. The current study compared its validity in discriminating bona-fide versus feigned (via experimental malingering paradigm) psychopathology against that of the Structured Inventory of Malingered Symptomatology (SIMS). Specifically, 452 Italian adult volunteers participated in this study: 216 were individuals with mental illness who were asked to take the SIMS and IOP-29 honestly, and 236 were nonclinical participants (experimental simulators) who took the same two tests with the instruction to feign a psychopathological condition. Two main, broad categories of symptom presentations were investigated: (a) psychotic spectrum disorders and (b) anxiety, depression, and/or trauma-related disorders. Data analysis compared the effect sizes of the differences between the patients and experimental simulators, as well as the AUC and classification accuracy statistics for both the SIMS and IOP-29. The results indicate that the IOP-29 outperformed the SIMS, with the differences between the two tools being more notable within the psychotic (IOP-29 vs. SIMS: d = -1.80 vs. d = -1.06; AUC = .89 vs. AUC = .79) than within the anxiety, depression, and/or trauma related subgroup (IOP-29 vs. SIMS: d = -2.02 vs. d = -1.62; AUC = .90 vs. AUC = .86). This study also demonstrates that the IOP-29, with its single cutoff score, is generalizable culturally and linguistically from the U.S. (English) to Italy (Italian).

Keywords: Inventory of Problems; SIMS; Malingering; Psychosis; Anxiety; Depression

A Clinical Comparison, Simulation Study Testing the Validity of SIMS and IOP-29 with an Italian Sample

In forensic evaluations, discriminating whether a given symptom presentation is bona-fide or alternatively non-credible is not an easy task. Currently, practitioners may rely on various methods, such as comprehensive interview-based measures, validity scales embedded in multiscale personality inventories and cognitive tests, and stand-alone, symptom validity (SVTs) and performance validity (PVTs) tests. However, the ultimate decision is often difficult.

A major issue is that there is always some uncertainty about which cutoff should be used in which situation, for any given measure assessing symptom validity. For example, for the Structured Inventory of Malingered Symptomatology (SIMS; Smith & Burger, 1997; Widows & Smith, 2005) – arguably one of the most widely utilized, stand-alone SVTs (Dandachi-FitzGerald, Ponds, & Merten, 2013; Martin, Schroeder, & Odland, 2015) – at least four different cutoffs (i.e., total score ≥15, ≥17, ≥20, or ≥25) have been suggested in the literature (e.g., van Impelen, Merckelbach, Jelicic, & Merten, 2014). Likewise, for the validity scales embedded in the second edition of the Minnesota Multiphasic Personality Inventory (MMPI–2; Green, 1991), the range of potentially optimal cut scores varies dramatically – for example, from ≥5 to ≥10 for the Fp, and from ≥9 to ≥31 for the F scale (Rogers, Sewell, Martin, & Vitacco, 2003). The same is true for the validity scales of the Personality Assessment Inventory (PAI; Hawes & Bocaccini, 2009; Morey, 1991, 2007).

Moreover, the great majority of currently available symptom validity measures perform differently across situations, contexts and symptom presentations. For example, the SIMS has a low specificity when testing individuals with schizophrenia or intellectual disability (van

Impelen et al., 2014). Also, the MMPI-2 Ds scale is known to perform sub-optimally when used with challenging symptom presentations involving PTSD or psychotic symptoms (Rogers et al., 2003). Similarly, the PAI validity scales tend to be more effective with severe and psychotic disorders than they are with mood or anxiety disorders (Hawes & Bocaccini, 2009). The Test of Memory Malingering (TOMM; Tombaugh, 1996) works better when used in the assessment of feigned cognitive impairment rather than with other symptom presentations. This context-dependent variability further complicates symptom validity assessment by the fact that the same score might mean different things for different people, with the evaluator not being able to tell the extent to which his or her testing results may be applicable to the specific person being tested (especially in the case of mixed symptoms presentations or comorbidities).

To overcome some of these generalizability and utility limitations, Viglione, Giromini, and Landis (2017) recently introduced the Inventory of Problems – 29 (IOP-29). Its goal is to provide practitioners with a brief, easy-to-use, decision-making tool that may be applied to a wide variety of contexts and clinical conditions, including neuropsychological impairment, psychosis, PTSD, and depression. Its key measure, the False Disorder Score (FDS), is a simple probability score that varies from zero to one. Scores approaching one suggest that the report is not credible, whereas scores approaching zero suggest that the report is valid. Thus, the FDS may be conceived of as an easy-to-use probability score with a stable cutoff that could add relevant information to the multidimensional assessment of symptom validity.

Viglione et al. (2017) presented multiple, scale development and independent, cross-validation samples with American adults. Using an a priori cutoff probability score of FDS $\geq$ .50, they reported sensitivity and specificity of about 80% with multiple, diverse, symptom presentation pictures. Importantly, despite its consisting of only 29 items, in Viglione et al.'s

(2017) validation studies the IOP-29 performed similarly to the longer and more complex MMPI-2 and PAI, while achieving better results than the TOMM. In fact, with a sample of 88 adult volunteers who were either genuinely experiencing (*n* = 43) or feigning via experimental malingering paradigm (*n* = 45) schizophrenic symptoms, the IOP-29 showed a receiver-operator characteristic curve (area under the curve; AUC) of .85, whereas the MMPI-2 scales F, Fp, and Ds-r2 showed AUC values of .79, .81, and .88 respectively. With an independent sample (*n* = 85) comprised of 43 individuals affected by depression and 42 simulators instructed to feign depression via experimental malingering paradigm, the IOP-29 showed an AUC of .90, whereas the TOMM showed AUC values of .76 (TOMM Trial 1), .82 (TOMM Trial 2) and .81 (TOMM Total). With a forensic sample including 128 volunteer offenders on community-based county and federal probation, half of which were asked to respond genuinely and the other half to simulate mental health symptoms via experimental malingering paradigm, the IOP-29 showed an AUC of .94. In contrast, the PAI scales NIM, MAL, and RDF showed AUC values of .97, .84, and .89 respectively. Lastly, with another independent sample of 90 adult volunteers, genuinely suffering (*n* = 45) or feigning psychotic symptoms via experimental malingering paradigm (*n* = 45), the IOP-29 showed an AUC of .92, whereas PAI scales NIM, MAL and RDS showed, respectively, AUC values of .89, .91, and .89.

Given the similarity of the findings observed across multiple, independent validation studies, in their IOP-29 developmental paper, Viglione et al. (2017) stated, “It is reasonable to anticipate that the IOP–29 will generalize well and perform similarly in different cultural contexts and future clinical comparison simulation studies” (p. 542). To date, however, this hypothesis has not been tested and reported in the literature. Indeed, although Italian, Brazilian-Portuguese, European-Portuguese, Chinese, Dutch, French, and German IOP-29 versions are

currently being investigated, no published studies have yet replicated Viglione et al.'s (2017) findings either in the United States or in other countries with other languages.

Aim of the Study

The current study aimed at testing, via a clinical-comparison, simulation study, the validity of the IOP-29 in discriminating bona-fide versus feigned (via experimental malingering paradigm) psychopathology in an Italian sample. We focused on two main, broad categories of psychopathology: (a) psychotic spectrum disorders and (b) non-psychotic, affect-related disorders. Indeed, as reviewed above (e.g., Rogers et al., 2003; van Impelen et al., 2014), available symptom validity measures tend to perform differently with these two categories of mental health problems. Malingered psychotic-like symptoms are frequent in criminal forensic evaluations because they function to minimize the perception of culpability and to reduce punishment. Malingered depression, anxiety, and/or trauma-related symptoms, the affect-related disorders mentioned previously, are common in civil forensic evaluations with the aim to procure compensation for psychological injury, workman's compensation, harassment, etc. (Bush et al., 2005; Bush, Heilbronner, & Ruff, 2014; Larrabee, 2003; Rogers, 2008).

Moreover, because the SIMS appears to be the most widely utilized, stand-alone SVT in Europe and North America (Dandachi-FitzGerald, Ponds, & Merten, 2013; Martin, Schroeder, & Odland, 2015), we compared the validity of the IOP-29 against that of the SIMS. Relevant to cost-benefit analysis and utility, it should be noted that the SIMS is more than two and a half times longer than the IOP-29 with 75 items versus the 29 items, respectively.

Also, by conducting this study in Italy, we tested the cross-cultural and linguistic generalizability of the IOP-29 findings. That is, we also evaluated whether the IOP-29 may be subject to cultural or language bias, as well as its validity in a non-U.S. country such as Italy.

Method

The data analyzed in this paper come from multiple research projects conducted by a multitude of Italian graduate students, in collaboration with several Italian psychiatric facilities. The main goal of each of these projects was to compare the classification accuracy of the IOP-29 against that of the SIMS. As such, all patients included in these studies were asked to answer the items of the tests honestly, whereas all nonclinical volunteers (experimental simulators) were asked to take the tests with the instruction to feign a psychopathological, mental health condition.

Participants

Patients. Patients were recruited at several psychiatric facilities, located in the north and central regions of Italy, after receiving approval by the pertinent ethical committees. In all cases, the diagnoses were made by psychiatrists at the referring sites according to the DSM-5 criteria. Psychiatrists also evaluated the presence of cognitive impairment (e.g., attentional problems) and other potential conditions that would not allow the patient to comply with the requirements of the research project. For example, if a patient was actively psychotic, he or she was not invited to take part in the research. Conversely, those who were judged to be able to comply with the research requirements were invited by the psychiatrist to take part in the study, and – if they showed some interest – they were put in contact with the graduate student leading the project at that specific site. Finally, only those patients who were able to read and sign the informed consent form in Italian were included.

The patient sample included 216 adults, with either a psychotic spectrum disorder (n = 89) or a non-psychotic, anxiety, depression, and/or trauma related disorder (n = 127). Within the psychotic spectrum disorder subgroup, slightly more than half were inpatients (55%). The most frequent diagnosis was schizophrenia (61.8%); in this group, however, we also included a

relatively large number of patients with schizoaffective disorder (10.1%), and a few cases of depression (6.7%) or bipolar disorder (3.4%) with psychotic features. Within the non-psychotic, anxiety, depression, and/or trauma related disorder subgroup, the percentage of inpatients was 48%. In this group, the most represented diagnosis was major depression (35.4%), followed by generalized anxiety disorder (17.3%), bipolar disorder in a depressive phase (12.6%), panic disorder (9.4%), obsessive-compulsive disorder (4.7%), and dysthymic disorder (2.4%).

Ages ranged from 18 to 81 (M = 49.4; SD = 14.0), and genders were relatively balanced, with 52.3% being women. All but nine patients (4.2%), were born in Italy, and these nine individuals moved to Italy during childhood. About half had never been married (49.1%), a third were married (29.6%), and the remaining ones were either divorced (7.9%), separated (7.4%), widowed (3.7%) or cohabitating (2.3%). The great majority had a high school degree or less (85.2%), with only few individuals having a bachelor's degree or more (14.8%).

Experimental simulators. After the university ethics committee approved the research project, experimental simulators were recruited via the snowball sampling method, and by posting advertisements on various social networks. Prospective participants were informed that they would be asked to take two psychological tests with the task to feign a mental disorder. Inclusion criteria required the following: being adult (i.e., age ≥ 18); being able to read and sign an informed consent form in Italian; and not being currently treated with psychiatric medication.

Experimental simulators were 236 adult volunteers, ranging in age from 18 to 77 (M = 32.9; SD = 12.7). Slightly more than half were women (56.2%), and all were born in Italy, except for one person, who had moved to Italy during childhood. The great majority had never been married (72.0%); a relatively small proportion was married (18.2%), and the remaining ones reported to be cohabitant (3.8%), divorced (1.7%), separated (1.3%), widowed (.8%), or other

(2.1%). In terms of education, 67.4% had a bachelor's degree or more, and 32.6% had a high school degree or less. All simulators but one self-reported that they had never been psychiatrically hospitalized for psychological reasons. This person clarified that he had been hospitalized long time ago, during adolescence. Furthermore, he was classified as a simulator both by the SIMS and by the IOP-29. Thus, including or excluding him from the participants of this study would not notably affect the main findings of our research.

Demographic Characterization of Patient vs. Simulator Samples. As shown in Table 1, the two samples did not differ significantly in terms of gender, but the simulator sample was significantly younger, more educated, and more frequently married or cohabitant. Accordingly, additional analyses were performed to examine the extent to which this heterogeneity could affect our main results (see below).

Procedures

Patients. First, all patient volunteers signed an informed consent form. Next, they were taken to a quiet room by the experimenter (the graduate student leading the project at that site) where they were administered the IOP-29 and SIMS in counterbalanced order. No compensation was given to the patients involved in this research. However, all were given the opportunity to learn more about the details and results of the study after completion of the tests, in case they were interested.

Simulators. After reading and signing the informed consent form, each participant was given some time to study the instructions, which included a vignette (i.e., a description of a real-life scenario in which an individual might be tempted to feign a mental disorder) to facilitate feigning (Rogers & Gillard, 2011; Viglione et al., 2001). In total, four vignettes were used in this study: one addressed feigning of psychotic symptoms, the other three addressed feigning of

anxiety, depression, and/or trauma related symptoms. The vignette to facilitate feigning of psychotic symptoms was taken from O'Brien (2004); the vignettes to facilitate feigning of depression, anxiety and/or trauma related symptoms were taken from McCullaugh (2011), Abramsky (2005), and Viglione et al. (2017). All were rich in contextual and historical detail.

As an incentive for careful and effortful participation, simulators were informed that the best simulators – i.e., those who would be able to look more "genuine" at both measures – would receive a small monetary compensation. In most cases, they were also given some time to search on the internet for some information about the psychopathology they were asked to feign. Furthermore, all were cautioned that if they presented their symptoms too dramatically, their performance would likely not be credible (Rogers & Bender, 2013; Viglione et al., 2001). To complete the study, all simulators completed the IOP-29 and SIMS in counterbalanced order in the roles depicted in the vignettes, i.e., as if they were trying to appear psychotic, anxious, depressed, or mentally disturbed within a real-life scenario.

Measures

All participants were administered both the SIMS and the IOP-29.

Structured Inventory of Malingered Symptomatology (SIMS; Smith & Burger, 1997; Widows & Smith, 2005). The SIMS contains 75 items describing implausible, rare, atypical, or extreme symptoms that bona-fide patients tend to not endorse. For the SIMS, the official, Italian adaptation translated by La Marca, Rigoni, Sartori and Lo Priore (2012) was used. All 75 items offer two response options: "true" and "false." Item responses may be grouped into five main scales, addressing the validity of symptoms related to Psychosis, Neurological Impairment, Amnestic Disorders, Low Intelligence, and Affective Disorders. The total number of implausible symptoms endorsed by the test-taker represents the Total Score,

which is the main, symptom validity scale of the SIMS. Indeed, the five SIMS subscales were not designed to detect over-reporting of mental health problems, but rather to learn what type of psychopathology the test-taker is over-reporting, when the SIMS Total Score is above the cutoff (Widows & Smith, 2005).

Although Smith and Burger (1997) originally recommended that the SIMS Total Score cutoff should be ≥15 (i.e., up to 14, the profile of the test-taker may be considered to be valid, if the Total Score is greater than 14 then the profile is invalid), another commonly employed cutoff is ≥17 (i.e., Total Score values greater than 16 are considered to be positive outcomes; van Impelen et al., 2014). Additionally, some authors have noted that using ≥15 or ≥17 as cutoffs generates too many false positive cases, and therefore they recommended to use a threshold of ≥20 (Clegg et al., 2009), or even ≥25 (Wisdom et al., 2010). A recent, meta-analytic study encompassing 4,180 protocols has provided support to the overall utility of the SIMS, although it also supported the claim that the specificity of the SIMS may be unsatisfactory when the traditional cutoffs (i.e., ≥ 15 and ≥ 17) are adopted (van Impelen et al., 2014).

Inventory of Problems – 29 (Viglione, Giromini, & Landis, 2017). The IOP-29 is comprised of 29 items addressing various mental health symptoms, test-related behaviors, attitudes towards one's own condition (e.g., ability to manage one's own symptoms, externalization of responsibility, denying any ability to influence one's condition, endorsing positive personal attributes, etc.) and problem-solving abilities. Among the 29 items, 26 could be characterized as SVT items. The other three involve solving simple mathematical or logical problems. Of note, many of the SVT items do not focus on specific symptoms of particular disorders, but rather on how the person lives or copes with his or her problems, or whether he or she would admit positive traits or successes. The aim in using many different item strategies was

to maximize incremental validity among the items and generalizability across different disorders and contexts. Thus, it was hoped that the variety of items making up the IOP-29 would reveal different, non-redundant information concerning the overall credibility of the test-taker's "global" presentation.

For this study, an Italian version of the IOP-29 was developed by following the classic, "back-translation" method (Brislin, 1980; Geisinger, 2003; Van de Vijver & Hambleton, 1996). That is, first, one of the developers of the IOP-29 translated the original, English IOP-29 into Italian; next, an Italian-American, bilingual psychologist who was blind to the original version of the IOP-29 back-translated this Italian version into English; lastly, two of the developers of the IOP-29 compared the two English versions (i.e., the original and the back-translated) to address inconsistencies and revise the translations as needed.

Twenty-seven IOP-29 items offer three response options: "true," "false," and "doesn't make sense" (26 of these items are the SVT items mentioned above, the other one of the three problem-solving items included in the IOP-29). The two remaining items are mathematical problems involving calculation and logic and require the respondent to write the correct numbers. The responses are then analyzed using a logistic regression-derived formula to generate the main feigning index of the IOP-29, the FDS, which is expressed as a probability score. The higher the FDS, the lower the credibility of the reported symptoms. As noted above, in their IOP-29 development research, Viglione et al. (2017) tested the IOP-29 with several forensic and clinical samples collected in the U.S. Using an a priori cutoff score of FDS $\geq$ .50, and with a base rate of .50, the IOP-29 demonstrated sensitivity, specificity, positive predictive power, negative predictive power, and overall correct classification values of about .80. Cohen's *d* effect sizes

representing the difference between honest respondents and experimental feigners were ≥ 1.7, and AUC values exceeded .84 in all cross-validation samples.

Data Analysis

The main purpose of this study was to test the validity of the IOP-29 and compare it against that of the SIMS. Accordingly, we inspected the effect sizes of the differences between the scores of patients and simulators, and examined AUC and classification accuracy of both the SIMS and IOP-29.

Results

As expected, patients and simulators significantly differed on both the SIMS and IOP-29 (Table 2). Among the SIMS scores, the biggest effect size was obtained by the Total Score in both the entire sample (d = -1.39) and within the psychotic spectrum subgroup (d = -1.06). Within the anxiety, depression, and/or trauma related subgroup, the biggest effect sizes in the SIMS were obtained by the Total Score (d = -1.62) and Low Intelligence scales (d = -1.65). As for the IOP-29, the FDS produced similar effect sizes across the three groups under consideration (i.e., psychotic subgroup, affect-related subgroup, and combined sample), and in all cases, d was ≤ -1.80.

Rogers et al. (2003) offered a scheme to classify the strength of Cohen's d effect sizes as "*moderate*" if $d \geq 0.75$; "*large*" if $d \geq 1.25$; "*very large*" if $d \geq 1.75$. Based on these benchmarks, the effect sizes produced by the SIMS Total Score may be characterized as *moderate* to *large*, depending on the sample under consideration (i.e., *moderate* for the psychotic subgroup, *large* for the anxiety, depression, and/or trauma related subgroup and the entire sample), whereas the effect sizes produced by the IOP-29 may be characterized as *very large* for all the subgroups and the entire sample.

Table 3 and Figure 1 present the AUC of the SIMS and IOP-29. The SIMS Total Score showed AUC values ranging from .79 (for the psychotic subgroup) to .86 (for the anxiety, depression, and/or trauma related subgroup), with AUC = .83 within the entire, combined sample. Conversely, the IOP-29 showed very similar statistics for all samples under investigation, with AUC values ranging from .89 (for the psychotic subgroup and the combined sample) to .90 (for the anxiety, depression, and/or trauma related subgroup).

To more directly compare the performance of the SIMS against that of the IOP-29, we calculated the point bi-serial correlations of SIMS and IOP-29 scores to group membership (dummy variable, with 0 = patient; 1 = simulator), and used procedures for comparing correlations described by Meng, Rosenthal, and Rubin (1992) to test whether the difference between these two correlations would be statistically significant. To obtain an effect size associated with these comparisons, we utilized the procedures described by Viglione, Giromini, Gustafson, and Meyer (2014) to convert Meng et al.'s (1992) *Z* statistic back to the *r* metric, and thus obtaining an *r* difference value (*r Diff*). As shown in Table 4, regardless of the sample under consideration, the IOP-29 significantly outperformed the SIMS, with the effect sizes of these correlation differences being *small* to *medium*, according to standard benchmarks interpreting *r* effect sizes (Cohen, 1988, 1992).

We also calculated classification accuracy of multiple and comparable, a-priori cutoffs for the SIMS and IOP-29 (Table 5). As noted above, four different cutoffs have been proposed for the SIMS. The three most commonly used cutoffs, i.e., ≥15, ≥17, and ≥20, tend to favor sensitivity over specificity; conversely, the less widely adopted cutoff of ≥25 seems to represent a better balance between sensitivity and specificity (van Impelen et al., 2014). Based on Viglione et al. (2017), with the IOP-29, cutoffs of ≥.20, ≥.30, and ≥.40 should favor sensitivity over

specificity similarly to the SIMS cutoffs of ≥15, ≥17, and ≥20. On the other hand, an IOP-29 cutoff of ≥.50 would balance sensitivity and specificity, similar to the SIMS cutoff of ≥25.

Additionally, because in forensic assessment of symptom validity, specificity has priority over sensitivity, given the high cost of false positive errors, we also identified SIMS and IOP-29 cutoffs that would yield, in our data, values of specificity as close as possible to .80, .90 and .95. Conversely, for screening purposes, sensitivity might be more important than specificity. Accordingly, we also identified SIMS and IOP-29 cutoffs that would yield sensitivity values as close as possible to .80, .90, and .95. The results of these "a-posteriori established" cutoff scores are reported in Table 5, too. Furthermore, given that positive predictive power (PPP) and negative predictive power (NPP) are highly dependent on the base rate of the condition being tested (Meehl & Rosen, 1955), and since the estimates of base rates for invalid profiles vary from one study to another (Young, 2015), Table 5 also presents PPP and NPP values calculated, via Streiner's (2003) formulas, for base rates of .10, .20, .30, .40, and .50.

In line with the hypothesis that the IOP-29 offers a better diagnostic accuracy than the SIMS, within the entire, combined sample the values of classification accuracy were higher for the IOP-29 than the SIMS for almost all a-priori cutoffs. For example, when looking at the highest a-priori cutoff values taken under consideration (i.e., ≥25 for the SIMS and ≥.50 for the IOP-29), overall correct classification (OCC) is .82 for the IOP-29 and .76 for the SIMS. For the second highest cutoff (i.e., ≥20 for the SIMS and ≥.40 for the IOP-29), OCC is .80 for the IOP-29 and .72 for the SIMS. With the second lowest cutoff (i.e., ≥17 for the SIMS and ≥.30 for the IOP-29), OCC is .75 for the IOP-29 and .69 for the SIMS. With the lowest a-priori cutoff value under investigation (i.e., ≥15 for the SIMS and ≥.20 for the IOP-29), however, these differences became much smaller, with OCC being .68 for the IOP-29 and .67 for the SIMS. Similar

considerations may be made when looking at the two subgroups too, although the advantage of the IOP-29 over the SIMS is greater for the psychotic spectrum subgroup, rather than for the anxiety, depression, and/or trauma related subgroup.

Table 5 results with a-posteriori cutoffs also support similar conclusions. Using cutoff scores that yield high levels of sensitivity generates greater specificity with the IOP-29 than with the SIMS. For example, within the entire sample, when sensitivity is about .80, .90, and .95, specificity is, respectively, .69, .48, and .17 for the SIMS and .82, .60, and .30 for the IOP-29. Likewise, when using cutoff scores that ensure high levels of specificity, sensitivity is greater for the IOP-29 than for the SIMS. Indeed, within the same group, i.e., the entire sample, when specificity is about .80, .90, and .95, sensitivity is, respectively, .75, .61, and .50 for the SIMS and .81, .73., and .65 for the IOP-29. Similar considerations may be drawn when looking at the different diagnostic subgroups and by looking at PPP and NPP values for different base rates.

Additional Analyses

Because the patient and simulator samples differed on age, marital status, and education, we performed additional analyses to examine the extent to which this heterogeneity could influence our results. More specifically, we removed the effects of each of those demographic variables, one at a time, through a series of partial correlations of SIMS and IOP-29 with group membership (dummy variable, 0 = patient; 1 = simulator). These partial correlations yielded nearly identical values to the point bi-serial correlations reported in Table 4, thus suggesting that the heterogeneity between the patient and simulator samples does not account for the superior performance of the IOP-29.

Discussion

Evaluating the credibility of psychopathological complaints plays a key role in forensic evaluations (Bush et al., 2005; Bush, Heilbronner, & Ruff, 2014; Merten, Merkelbach, Giger, & Stevens, 2016). In this study, we evaluated the detection efficacy of a widely adopted, stand-alone, SVT, i.e., the SIMS, against that of a newly developed measure, i.e., the IOP-29. Although both instruments demonstrated validity in our clinical comparison, simulation study conducted in Italy, the IOP-29 outperformed the SIMS, with the differences between the two tools being more notable within the psychotic than within the anxiety, depression, and/or trauma related subgroup.

In a recent meta-analysis of the SIMS, van Impelen et al. (2014) reported that across the 24 simulation studies examined in the literature, the SIMS differentiated between experimental feigners and honest responders with a Cohen's *d* effect size ranging from 0.5 to 4.7. However, the authors also highlighted that some of the studies included in their meta-analysis employed nonclinical adults, rather than patients, as controls, and that "nonclinical controls tend to produce significantly lower SIMS scores than patient controls" (van Impelen et al., 2014, p. 1343). As such, the effect sizes obtained in our study by the SIMS Total Score, ranging from $d = 1.06$ to $d = 1.62$, appear comparable to previous findings obtained in the literature with clinical comparison, simulation studies. Thus, in our study the SIMS demonstrated validity in discriminating between bona-fide versus experimentally feigned psychopathological complaints.

The IOP-29, however, consistently demonstrated greater validity than the SIMS when considering a wide variety of base rates and both sensitivity and specificity benchmarks. Not only were its AUC and effect sizes larger than those obtained by the SIMS, the IOP-29 FDS performed similarly well with the psychotic and anxiety, depression, and/or trauma related subgroups. Conversely, validity and optimal cutoffs for the SIM differed across conditions and disorders. As noted in the Introduction, the fact that the same measure is differently valid for

different conditions complicates symptom validity interpretation and assessment. For example, a practitioner might not know the extent to which his or her testing results may be applicable to a case with a mixed symptoms presentation. Thus, the IOP-29 might offer an important advantage, compared to the SIMS and other instruments from a practical or utility standpoint. Potentially, it might indeed be used with many, various, and different cases, without adjusting the cutoff scores and interpretative inferences across different cases, contexts, and diagnoses.

It is also noteworthy that the IOP-29, despite its consisting of 29 items only, performed significantly better than the 75-item long SIMS. In current practice, the SIMS is typically used as a screening tool, to decide whether follow-up testing is advised (Widows & Smith, 2005). To this extent, if future studies confirmed the results of the current study, the IOP-29 might be preferable to the SIMS, as it likely requires a shorter administration time. It should be noted that when using a SVT as a global screening tool, sensitivity might be more important than specificity. Moreover, Table 3 and Figure 1 reveal that when emphasizing sensitivity by using low, a-priori cutoff scores, the differences between the IOP-29 and SIMS are less evident. Furthermore, a brief version of the SIMS has recently been developed (Malcore, Schutte, Van Dyke, & Axelrod, 2015), so that future research comparing the IOP-29 against the shorter SIMS should be considered.

Because the IOP-29 has been developed and validated throughout 14 independent studies encompassing thousands of patients and simulators, Viglione et al. (2017) anticipated that the shrinkage from their developmental research to future, cross-validation studies would be minimal. Our study confirms this expectation. Indeed, in their IOP-29 development paper, Viglione et al. (2017) stated, “with base rates of .5 [the IOP-29] produced sensitivity, specificity, positive predictive power, and negative predictive power statistics of about .80” (p. 534). As seen

in Table 5, with the same cutoff score of FDS ≥ .50 and base rate of .50, these same classification accuracy statistics surpassed .80 in both the entire sample and the psychotic and anxiety, depression, and/or trauma related subgroups. Under the same condition, these same classification accuracy statistics surpassed .80 in only three out of twelve data points for the medium, a-priori SIMS cutoff (≥25) and BR of .50. Given the replication of essentially no shrinkage in our study, it may be speculated that future clinical comparison, simulation studies will continue to achieve similar results as well.

Three additional remarks deserve mention. This is the first study testing the efficacy of the IOP-29 in a non-English speaking population. Accordingly, the IOP-29 is demonstrating validity and utility across culture, language, and geography. Secondly, Viglione et al. (2017) did not administer the IOP-29 in its final, paper-and pencil format used in this study. Instead, the IOP-29 items were interspersed in a 181 item, computerized format. Thus, our study is the first to examine the validity of the IOP-29 in its final, paper-and-pencil format. Thirdly, this study supports the utility of the IOP-29 to practitioners in forensic/legal applications. Indeed, the FDS has been cross-validated so far in six U.S. samples (n = 549) and one Italian sample (n = 452), encompassing 1001 participants (i.e., patients: n = 427; experimental simulators: n = 446; forensic sample: n = 128). It is worth noting that the IOP-29, using the same cutoff scores, reached similar results in all these samples encompassing psychotic, depression, anxiety, post-traumatic and cognitive diagnoses. The current study specifically fortifies the generalizability and stability of the findings presented by Viglione et al. (2017) across culture, context, diagnosis, and test format.

Nonetheless, it should be noted the IOP-29 still needs to be tested with real-life situations, with a known-group research design. Like all simulation studies, indeed, the current

investigation cannot rule out that real life malingerers will behave differently from our experimental simulators, as feigning a mental disorder in an experimental context may be different from malingering in a real-life context. Thus, clinical studies with multiple symptom validity scales and indices used as criterion measures are needed to further validate the ecological validity of the IOP-29. Also, generalization to cognitive and neuropsychological disorders has limited research support so far, so that both simulation and known group studies with this population are needed.

Additional limitations include the fact that our patient and experimental simulator groups were not well counterbalanced in terms of age, marital status, and education. Our additional analyses provide some support for discounting the effect of these demographic variables, but future studies with more homogenous contrast groups are needed. As for the participant education levels, the only information available for this sample is the maximum degree obtained by the participants, so that it is not clear whether the participants who have a high school degree might have also attended few years of college. Furthermore, although we instructed our investigators to counterbalance the order with which the two tests were administered, we did not record the administration order for each participant, so order effects could not evaluated. Future research could overcome this limitation. Likewise, we did not assess reading level, so that future studies might also consider evaluating the effects of limited literacy on IOP-29 and SIMS performance. Finally, during the translation and back-translation procedure, we did not administer both the back translated version and the original, English version to non-Italian, English as a first language individuals to test potential differences in the IOP-29 scores. Given the need to translate the IOP-29 into other languages to further evaluate its cross-cultural

applicability, future studies might administer the back-translated and the original versions of the IOP-29 to a sample of English speaking individuals.

Despite these limitations, in this Italian sample our study provides important support for both the cross-cultural generalizability, validity, and utility of the newly developed and brief IOP-29 and also evidence of its superiority over the SIMS. Future research could compare the classification accuracy of the IOP-29 against other PVTs and SVTs regularly used in clinical and forensic practice. For example, it might be particularly useful to administer the IOP-29 together with a word memory based PVT using the forced choice recognition paradigm such as the Word Choice Test, to test modality specificity as a potential confound (Erdodi, 2017), or with a quick-and-easy, hard-to-fail PVT like the Rey 15-item test.

Compliance with Ethical Standards

Conflict of Interest: When this study was designed and realized, none of the authors had any conflict of interest. However, the first and second authors are currently in the process of creating a Limited Liability Company for the commercial use of the IOP-29. Conversely, the third and fourth authors continue to have no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

Abramsky, A. B. (2005). *Assessment of test behaviors as a unique construct in the evaluation of malingered depression on the Inventory of Problems: Do test behaviors add significant variance beyond problem endorsement strategies?* (Unpublished doctoral dissertation). California School of Professional Psychology, San Diego, CA.

Berthelson, L., Mulchan, S. S., Odland, A. P., Miller, L. J., & Mittenberg, W. (2013). False positive diagnosis of malingering due to the use of multiple effort tests. *Brain Injury, 27*, 909–916.

Brislin, R. W. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 1, pp. 389-444). Boston, MA: Allyn & Bacon.

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*, 197–205.

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R., & Silver, C. H. (National Academy of Neuropsychology Policy & Planning Committee) (2005). Symptom validity assessment: Practice issues and medical necessity. Official position of the National Academy of Neuropsychology. *Archives of Clinical Neuropsychology, 20*, 419–426.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Dandachi-FitzGerald, B., Ponds, R. W. H. M., & 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*, 771–783.

Davis, J. J., & Millis, S. R. (2014). Examination of performance validity test failure in relation to number of tests administered. *The Clinical Neuropsychologist, 28*, 199–214.

Erdodi, L. A. (2017). Aggregating validity indicators: The salience of domain specificity and the indeterminate range in multivariate models of performance validity assessment. *Applied Neuropsychology: Adult [Epub ahead of print, Nov 7]*. Doi: 10.1080/23279095.2017.1384925

Geisinger, K. F. (2003). Testing and assessment in cross-cultural psychology. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology. Part one* (pp. 95–117). Hoboken, NJ: John Wiley & Sons, Inc.

Giger, P., Merten, T., Merckelbach, H., & Oswald, M. (2010). Detection of feigned crime-related amnesia: A multi-method approach. *Journal of Forensic Psychology Practice, 10*, 140-163.

Green, R. L. (1991). *MMPI-2/MMPI: An interpretative manual*. Boston, MA: Allyn & Bacon.

Hawes, S., & Bocaccini, M. (2009). Detection of overreporting of psychopathology on the Personality Assessment Inventory: A meta-analytic review. *Psychological Assessment, 21*, 112–124.

La Marca, S., Rigoni, D., Sartori, G., & Lo Priore, C. (2012). *Structured Inventory of Malingered Symptomatology (SIMS): Manuale* (2nd ed.). Firenze, IT: Giunti O.S.

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist, 17*, 54–68.

Larrabee, G. J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology, 29*, 364–373.

Malcore, S. A., Schutte, C., Van Dyke, S. A., & Axelrod, B. N. (2015). The development of a reduced-item Structured Inventory of Malingered Symptomatology (SIMS). *Psychological Injury and Law, 8*, 95–99.

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey on North American professionals. *The Clinical Neuropsychologist, 29*, 741–776.

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

Meehl, P., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychological bulletin, 52(3), 194-216.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172–175.

Merten, T., Merkelbach, 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*, 102-111.

Morey, L. C. (1991). *Personality Assessment Inventory. Professional manual*. Odessa, FL: Psychological Assessment Resources.

Morey, L. C. (2007). *Personality Assessment Inventory (PAI). Professional manual* (2nd ed.). Odessa, FL: Psychological Assessment Resources.

O'Brien, S. M. (2004). *An investigation into the incremental value of test dependent malingering of schizophrenia* (Unpublished doctoral dissertation). California School of Professional Psychology, San Diego, CA.

Rogers, R. (2008). Detection strategies for malingering and defensiveness. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 14–35). New York, NY: Guilford Press.

Rogers, R., & Bender, S. D. (2013). Evaluation of malingering and related response styles. In I. B. Weiner (Ed.-in-Chief), J. R. Graham, & J. A. Naglieri (Vol. Eds.), *Comprehensive handbook of psychology: Assessment psychology* (2nd ed., Vol. 10, pp. 517–540). Hoboken, NJ: Wiley.

Rogers, R., & Gillard, N. D. (2011). Research methods for the assessment of malingering. In B. Rosenfeld & S. Penrod (Eds.), *Research methods in forensic psychology* (pp. 174–188). Hoboken, NJ: Wiley.

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*, 160–177.

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.

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

Van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*, 89-99.

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*, 1336–1365.

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*, 297–313.

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*, 534-544.

Viglione, D. J., Wright, D., Dizon, N. T., Moynihan, J. E., DuPuis, S., & Pizitz, T. D. (2001). Evading detection on the MMPI–2: Does caution produce more realistic patterns of responding? *Assessment, 8*, 237–250.

Viglione, D., Giromini, L., Gustafson, M. L., & Meyer, G. J. (2014). Developing continuous variable composites for Rorschach measures of thought problems, vigilance, and suicide risk. *Assessment, 21*, 42-49.

Widows, M. R., & Smith, G. P. (2005). *SIMS-Structured Inventory of Malingered Symptomatology. Professional manual*. Lutz, FL: Psychological Assessment Resources.

Young, G. (2015). Malingering in Forensic Disability-Related Assessments: Prevalence 15±15 %. *Psychological Injury and Law, 8(3)*, 188–199.

Table 1

Demographic Composition of the Patient and Simulator Samples

	Patients (n = 216)	Simulators (n = 236)
Age [$t(434.1) = 13.10$, $p < .01$, $d = 1.24$]		
M	49.4	32.9
SD	14.0	12.7
Gender [*Phi* = .04, $p = .41$]		
Men	103 (47.7%)	103 (43.8%)
Women	113 (52.3%)	132 (56.2%)
Marital status [*Phi* = .13, $p < .01$]		
Married/Cohabitant	69 (31.9%)	52 (22.0%)
Not married/Not cohabitant	147 (48.1%)	184 (78.0%)
Education [*Phi* = .53, $p < .01$]		
Bachelor's degree or more	32 (14.8%)	159 (67.4%)
High school degree or less	184 (85.2%)	77 (32.6%)

Note. One case was missing gender information.

Table 2

Comparison between Patients and Simulators on SIMS and IOP-29 Mean Scores

	Patients		Simulators		*t*	*p*	*d*
	M	*SD*	*M*	*SD*			
Entire Sample (N = 452)							
SIMS							
Psychosis	2.8	2.9	7.5	4.6	-13.1 [a]	<.01	-1.21
Neurological Impairment	4.0	3.0	7.2	3.7	-10.0 [a]	<.01	-0.94
Amnestic Disorders	2.8	2.6	6.8	4.0	-12.6 [a]	<.01	-1.16
Low Intelligence	3.2	1.8	6.2	3.6	-11.6 [a]	<.01	-1.07
Affective Disorders	6.7	2.4	7.6	2.7	-3.7 [a]	<.01	-0.35
Total Score	19.5	9.3	35.3	13.0	-14.9 [a]	<.01	-1.39
IOP-29							
False Disorder Score	.30	.19	.73	.25	-20.7 [a]	<.01	-1.93
Psychosis Subgroup (n = 214)							
SIMS							
Psychosis	4.5	3.2	8.0	4.3	-6.7 [a]	<.01	-0.89
Neurological Impairment	4.9	3.1	7.5	3.8	-5.4	<.01	-0.74
Amnestic Disorders	3.5	2.8	6.5	3.8	-6.5 [a]	<.01	-0.86
Low Intelligence	3.5	1.8	4.8	2.7	-4.0 [a]	<.01	-0.53
Affective Disorders	6.2	2.3	8.3	2.5	-6.2	<.01	-0.86
Total Score	22.7	10.1	35.0	12.6	-7.7	<.01	-1.06
IOP-29							
False Disorder Score	.30	.20	.71	.25	-13.0	<.01	-1.80
Anx/Dep/Trauma Subgroup (n = 238)							
SIMS							

	Patients		Simulators		*t*	*p*	*d*
	M	*SD*	*M*	*SD*			
Psychosis	1.6	1.9	6.9	4.9	-10.9 [a]	<.01	-1.41
Neurological Impairment	3.4	2.8	6.8	3.6	-8.1 [a]	<.01	-1.05
Amnestic Disorders	2.3	2.4	7.1	4.2	-10.7 [a]	<.01	-1.39
Low Intelligence	2.9	1.7	7.9	3.8	-12.7 [a]	<.01	-1.65
Affective Disorders	7.1	2.3	6.9	2.7	.8 [a]	.44	0.10
Total Score	17.3	8.0	35.5	13.5	-12.5 [a]	<.01	-1.62
IOP-29							
False Disorder Score	.30	.19	.75	.25	-15.6 [a]	<.01	-2.02

Note. [a] Because homoscedasticity could not be assumed, Welch–Satterthwaite method was used to adjust degrees of freedom.

Table 3

Receiver Operator Characteristic Curve of SIMS and IOP-29

	AUC	SE
Entire Sample (N = 452)		
SIMS		
Psychosis	.79	.02
Neurological Impairment	.75	.02
Amnestic Disorders	.79	.02
Low Intelligence	.76	.02
Affective Disorders	.60	.03
Total Score	.83	.02
IOP-29		
False Disorder Score	.89	.02
Psychosis Subgroup (n = 214)		
SIMS		
Psychosis	.73	.03
Neurological Impairment	.71	.04
Amnestic Disorders	.73	.03
Low Intelligence	.64	.04
Affective Disorders	.73	.03
Total Score	.79	.03
IOP-29		
False Disorder Score	.89	.02
Anx/Dep/Trauma Subgroup (n = 238)		
SIMS		
Psychosis	.83	.03
Neurological Impairment	.77	.03
Amnestic Disorders	.83	.03
Low Intelligence	.86	.02
Affective Disorders	.46	.04
Total Score	.86	.02
IOP-29		
False Disorder Score	.90	.02

Note. AUC = area under the curve; SE = standard error.

Table 4

Point Bi-serial Correlations of SIMS and IOP-29 to Group Membership (0 = Patient; 1 = Simulator)

	SIMS Total	IOP-29 – FDS	*Z*	*p*	*r* Diff.
Entire Sample (*N* = 452)	.57	.69	-4.57	<.01	.21
Psychosis Subgroup (*n* = 214)	.47	.67	-4.91	<.01	.29
Anx/Dep/Trauma Subgroup (*n* = 238)	.64	.72	-2.05	.04	.14

Note. r Diff = *r* difference. All correlations significant at p < .01; FDS = False Disorder Score.

Table 5. Classification Accuracy of SIMS (on the left of the slash) *versus* IOP-29 (on the right of the slash)

	Cutoff	Se	Sp	BR = .10		BR = .20		BR = .30		BR = .40		BR = .50	
				PPP	NPP	PPP	NPP	PPP	NPP	PPP	NPP	PPP	NPP
Entire Sample													
A-priori cutoffs													
Very low	≥15 / ≥.20	.93 / .95	.41 / .40	.15 / .15	.98 / .98	.28 / .28	.96 / .97	.40 / .40	.93 / .94	.51 / .51	.89 / .92	.61 / .61	.85 / .88
Low	≥17 / ≥.30	.90 / .90	.48 / .60	.16 / .20	.98 / .98	.30 / .36	.95 / .96	.42 / .49	.92 / .93	.53 / .60	.88 / .90	.63 / .69	.82 / .86
Low to medium	≥20 / ≥.40	.85 / .89	.58 / .71	.18 / .26	.97 / .98	.34 / .44	.94 / .96	.46 / .57	.90 / .94	.57 / .67	.85 / .91	.67 / .76	.80 / .87
Medium	≥25 / ≥.50	.78 / .81	.73 / .82	.24 / .33	.97 / .98	.42 / .53	.93 / .95	.55 / .66	.88 / .91	.66 / .75	.83 / .87	.74 / .82	.77 / .81
A-posteriori cutoffs													
Se = .80	≥23 / ≥.51	.81 / 81	.69 / .82	.23 / .33	.97 / .98	.40 / .53	.94 / .95	.53 / .66	.90 / .91	.64 / .75	.85 / .87	.73 / .82	.79 / .82
Se = .90	≥17 / ≥.31	.90 / .90	.48 / .60	.16 / .20	.98 / .98	.30 / .36	.95 / .96	.42 / .49	.92 / .94	.53 / .60	.88 / .90	.63 / .69	.82 / .86
Se = .95	≥12 / ≥.16	.95 / .95	.17 / .30	.11 / .13	.97 / .98	.22 / .25	.94 / .96	.33 / .37	.90 / .94	.43 / .48	.85 / .91	.54 / .58	.79 / .87
Sp = .80	≥28 / ≥.51	.75 / .81	.81 / .82	.30 / .33	.97 / .98	.49 / .53	.93 / .95	.62 / .66	.88 / .91	.72 / .75	.83 / .87	.79 / .82	.76 / .82
Sp = .90	≥33 / ≥.64	.61 / .73	.90 / .93	.41 / .52	.95 / .97	.61 / .71	.90 / .93	.73 / .81	.84 / .89	.81 / .87	.78 / .84	.86 / .91	.70 / .77
Sp = .95	≥37 / ≥.70	.50 / .65	.95 / .97	.54 / .72	.95 / .96	.73 / .86	.88 / .92	.82 / .91	.82 / .87	.88 / .94	.74 / .81	.92 / .96	.65 / .74
Psychosis Subgroup													
A-priori cutoffs													
Very low	≥15 / ≥.20	.92 / .94	.24 / .40	.12 / .15	.96 / .98	.23 / .28	.92 / .96	.34 / .40	.87 / .94	.45 / .51	.82 / .90	.55 / .61	.75 / .86
Low	≥17 / ≥.30	.90 / .89	.34 / .60	.13 / .20	.97 / .98	.25 / .35	.93 / .96	.37 / .49	.88 / .93	.47 / .59	.83 / .89	.58 / .69	.76 / .84
Low to medium	≥20 / ≥.40	.88 / .88	.43 / .65	.15 / .22	.97 / .98	.28 / .39	.93 / .96	.40 / .52	.89 / .93	.51 / .63	.84 / .89	.61 / .72	.78 / .84
Medium	≥25 / ≥.50	.83 / 82	.61 / .81	.19 / .32	.97 / .98	.35 / .52	.94 / .95	.48 / .65	.89 / .91	.59 / .74	.84 / .87	.68 / .81	.78 / .81
A-posteriori cutoffs													
Se = .80	≥27 / ≥.50	.80 / .82	.63 / .81	.19 / .32	.97 / .98	.35 / .52	.93 / .95	.48 / .65	.88 / .91	.59 / .74	.83 / .87	.68 / .81	.76 / .81
Se = .90	≥17 / ≥.30	.90 / .89	.34 / .60	.13 / .20	.97 / .98	.25 / .35	.93 / .96	.37 / .49	.88 / .93	.47 / .59	.83 / .89	.58 / .69	.76 / .84
Se = .95	≥13 / ≥.12	.94 / .95	.16 / .20	.11 / .12	.96 / .97	.22 / .23	.92 / .94	.32 / .34	.87 / .91	.43 / .44	.81 / .86	.53 / .54	.74 / .81
Sp = .80	≥31 / ≥.50	.72 / .82	.80 / .81	.28 / .32	.96 / .98	.47 / .52	.92 / .95	.60 / .65	.87 / .91	.70 / .74	.81 / .87	.78 / .81	.74 / .81
Sp = .90	≥36 / ≥.57	.50 / .76	.91 / .87	.38 / .39	.94 / .97	.58 / .59	.88 / .94	.71 / .71	.81 / .89	.79 / .79	.73 / .84	.85 / .85	.65 / .78
Sp = .95	≥42 / ≥.64	.33 / .72	.94 / .93	.37 / .54	.93 / .97	.57 / .73	.85 / .93	.69 / .82	.77 / .89	.78 / .88	.68 / .83	.84 / .91	.59 / .77
Anx/Dep/Trauma Subgroup													
A-priori cutoffs													
Very low	≥15 / ≥.20	.94 / .96	.53 / .39	.18 / .15	.99 / .99	.33 / .28	.97 / .97	.46 / .40	.95 / .95	.57 / .51	.93 / .93	.66 / .61	.89 / .90
Low	≥17 / ≥.30	.90 / .92	.58 / .60	.19 / .20	.98 / .99	.35 / .36	.96 / .97	.48 / .50	.93 / .95	.59 / .60	.90 / .92	.68 / .70	.85 / .88
Low to medium	≥20 / ≥.40	.82 / .90	.69 / .76	.22 / .29	.97 / .99	.39 / .48	.94 / .97	.53 / .61	.90 / .95	.63 / .71	.85 / .92	.72 / .78	.79 / .88
Medium	≥25 / ≥.50	.71 / .81	.82 / .83	.30 / .34	.96 / .98	.50 / .54	.92 / .95	.63 / .67	.87 / .91	.72 / .76	.81 / .87	.80 / .82	.74 / .81

A-posteriori cutoffs														
Se = .80	≥20 / ≥.50	.82 / .81	.69 / .83	.22 / .34	.97 / .98	.39 / .54	.94 / .95	.53 / .67	.90 / .91	.63 / .76	.85 / .87	.72 / .82	.79 / .81
Se = .90	≥17 / ≥.37	.90 / .90	.58 / .76	.19 / .29	.98 / .99	.35 / .48	.96 / .97	.48 / .61	.93 / .95	.59 / .71	.90 / .92	.68 / .79	.85 / .88
Se = .95	≥12 / ≥.20	.96 / .96	.21 / .39	.12 / .15	.98 / .99	.23 / .28	.96 / .97	.34 / .40	.93 / .95	.45 / .51	.90 / .93	.55 / .61	.86 / .90
Sp = .80	≥24 / ≥.43	.75 / .87	.80 / .80	.30 / .32	.97 / .98	.49 / .51	.93 / .96	.62 / .64	.88 / .93	.72 / .74	.83 / .90	.79 / .81	.76 / .86
Sp = .90	≥29 / ≥.64	.69 / .73	.90 / .92	.43 / .51	.96 / .97	.63 / .70	.92 / .93	.74 / .80	.87 / .89	.82 / .86	.82 / .84	.87 / .90	.75 / .77
Sp = .95	≥33 / ≥.70	.61 / .68	.95 / .96	.59 / .66	.96 / .96	.76 / .81	.91 / .92	.85 / .68	.85 / .96	.90 / .92	.79 / .82	.93 / .95	.71 / .75

Note. Sample sizes are: N = 452 for the entire sample; n = 214 for the psychotic spectrum subgroup; and n = 238 for the anxiety, depression, and/or trauma related subgroup. Se = sensitivity; Sp = specificity; BR = base rate; PPP = positive predictive power; NPP = negative predictive power.

Figure 1. Graphical Representation of Receiver Operator Characteristic Curve of SIMS Total and IOP-29 False Disorder Score (FDS)