Assessing Negative Response Bias with the Inventory of Problems-29 (IOP-29): a Quantitative Literature Review

Assessing Negative Response Bias with the Inventory of Problems – 29 (IOP-29):

A Quantitative Literature Review

Abstract

This article reviews published, journal articles informing on the conditions of use, strengths, weaknesses, and optimal cut scores of the Inventory of Problems – 29 (IOP-29; Viglione & Giromini, 2020). To provide more accurate information on the convergent and incremental validity, hit rates, and optimal cut scores of the IOP-29, in addition to reviewing all published IOP-29 studies, we also retrieved all data sets associated with each of those studies, and performed some additional analyses. Taken together, the findings presented in this quantitative literature review indicate that: (a) the IOP-29 correlates more strongly with other symptom validity tests (SVTs) than with other performance validity tests (PVTs); (b) the IOP-29 yields incremental validity when used together with other validity checks; (c) its classification accuracy compares favorably to that of other established tools; (d) its suggested cut scores perform similarly well across various diagnoses and contexts. When considering the 3,777 IOP-29 protocols included in the statistical analyses comparing credible (k = 16) versus noncredible (k = 17) presentations, the weighted mean Cohen's d was 3.02 (weighted SD = .98; *range*: 1.48 – 5.31) and the weighted mean AUC was .95 (weighted SD = .04; *range*: .83 – 1.00). The standard IOP-29 cut score of ≥ .50 yielded a weighted mean sensitivity of .86 (*weighted SD* = .07; *range*: .63 – .96) at a weighted mean specificity of .92 (*weighted SD* = .06; *range*: .79 – 1.00). These statistics, however, could be inflated by the fact that almost all of examined studies used a simulation research paradigm.

Keywords: Inventory of Problems; IOP-29; malingering; review; symptom validity.

Assessing Negative Response Bias with the Inventory of Problems – 29 (IOP-29): A Quantitative Literature Review

This article, prepared for a Special Issue of *Psychological Injury and Law* on self-report measures of negative response bias (Giromini, Young, & Sellbom, 2022), reviews published journal articles informing on the conditions of use, strengths, weaknesses, and optimal cut scores of the Inventory of Problems – 29 (IOP-29; Viglione & Giromini, 2020). Although the IOP-29 is a relatively new instrument, which was first published in 2017 (Viglione, Giromini, & Landis, 2017) and officially released only in 2018 (www.iop-test.com), we were able to identify 15 research articles that investigated, or provided information about, its effectiveness and psychometric properties. These empirical investigations had been conducted in several different cultural contexts (US, UK, Australia, Italy, France, Portugal, etc.) and languages, with various research designs (e.g., between-subjects and repeated-measures simulation/analogue designs, known-groups designs, etc.) and a variety of diagnostic conditions and symptom presentations (e.g., depression, post-traumatic stress disorder, schizophrenia, and mild traumatic brain injury).

To identify relevant articles, we searched the term "IOP-29" in three databases, i.e., PsycINFO, PubMed, and Scopus (this search was conducted on November 3, 2021). Because the publication describing the development and initial validation of the IOP-29 occurred in 2017 (Viglione et al., 2017), we limited our search to years 2017 to 2021. Additionally, we only considered publications written in English language and published in peer-reviewed journals. Using these criteria, we identified a total of 69 articles (15 from PsycINFO, 12 from PubMed, 42 from Scopus), of which 25 were duplicates (Figure 1). Of the 44 unique sources identified, 7 were excluded after screening their abstracts. Next, of the 37 full-text records assessed for

inclusion, 19 were excluded because they did not administer the IOP-29 and three (Giromini et al., 2020c, 2021; Viglione et al., 2019) because they re-analyzed the same data sets utilized for other articles. As such, a total of 15 IOP-29 articles were included in this quantitative review, which is the first systematic effort to summarize the information and implications from this rapidly accumulating empirical research.

Background and Conditions of Use

The IOP-29 is a 29-item, self-administered, symptom validity test (SVT) designed to discriminate credible from noncredible presentations related to psychotic (e.g., Winters et al., 2020), depressive (e.g., Ilgunaite et al., 2020), cognitive/neuropsychological (e.g., Gegner et al., 2021), and trauma-related (e.g., Carvalho et al., 2021) problems, and combinations thereof (Viglione & Giromini, 2020). It was developed over a long period through an iterative item development and selection procedure. Initially, 27 potential feigning strategies or item groups with a total of 245 items were created, based on the clinical and forensic experience of the IOP authors, and on the accumulating research on malingering-related phenomena. Based on the results obtained with some initial, empirical studies, the items and response structures that best discriminated between bona fide patients and experimental feigners were retained; those that did not work well were eliminated or modified and a few additional items were developed. This resulted in a second set of 165 items to be further investigated. Noteworthy, our pilot studies had suggested that the keyed-true/answered-true test items, i.e., items describing symptoms or impairments which are answered true by feigners, shared among them a large amount of method variance that would decrease their unique prediction. Accordingly, this second set of items minimized the number of these *problem affirmation* or *pseudo-symptoms* items, which instead characterize the majority of currently available SVTs.

Some additional research efforts led to the development of various refinements of the instrument (e.g., some new items were generated, others were eliminated, some others were reformulated, etc.) and culminated with the identification of 181 items that were deemed to discriminate credible from noncredible presentations of psychotic, depressive, neuropsychological, and/or post-traumatic stress disorders with high accuracy and precision (Viglione et al., 2019). The 29 items of the IOP-29 were selected from these 181 items[1] as the best combination of items that, together, would best discriminate from credible and noncredible presentations and best generalize their validity from one condition or symptom presentation to another (Viglione et al., 2017). Item selection favored incremental validity of each item based on each item's unique contribution to the prediction of feigning of various symptom presentations. More specifically, the first selected item was the one that provided the best discrimination between credible and noncredible presentations of various conditions (PTSD, depression, psychosis, and neuropsychological impairment). Next, the following items were selected one by one, based on the amount of incremental validity each of them yielded after partialing out previously selected item(s): at each iteration, the item that would provide the greater amount of incremental validity was selected. This procedure was stopped after 29 iterations, when the next iteration would not yield any incremental validity over the previously selected 29 items. Thus, the IOP-29 is the best *set* of 29 items. These are not the 29 items that individually have the greatest predictive power. This is because, as a behavior, feigning is understood as a behavior in a context (Rogers & Bender, 2018), not as a singular construct or trait.

Different from most SVTs, which present the test-taker with a series of rare, very rare, or non-existing symptoms and identify noncredible presentations based on the frequency with

[1] These 29 items were also included in the previous 165-item version of the test.

which these pseudo-symptoms are endorsed, the IOP-29 uses multiple detection strategies. Indeed, in addition to the classic unlikely and amplified mental disorder detection strategies (Rogers, 2018), the items of the IOP-29 also address characteristics and behaviors such as (a) the externalization or minimization of the evaluee's responsibility concerning their psychological condition, (b) criticisms of and dramatic reactions to the evaluation context, (c) the refusal to admit qualified positive attributes or experiences, which patients and healthy controls almost always endorse, and (d) the presentation of specific impairments or difficulties that are particularly relevant to forensic and high stakes evaluations. These additional strategies emerged from a review of the literature on the effectiveness of interview techniques, which was conducted in the early development stages of the test (Rogers, 2008), as well as from the forensic experience accumulated by the second author of this article. More generally, rather than inquiring on whether or not the test-taker is experiencing a given list of symptoms, the IOP-29 evaluates how an individual copes with various psychological and/or neuropsychological problems in the particular evaluation context. Arguably, the IOP-29 approach to investigating the credibility of presented psychological problems would thus yield a kind of information that should not be overly redundant with that of the typical SVT. That is, the IOP-29 could offer a description of the patient/evaluee from a slightly different angle or perspective, compared to other SVTs like the Structured Inventory of Malingered Symptoms (SIMS; Smith & Burger, 1997) or the F scales of the Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher et al., 2001; MMPI–RF; Ben-Porath & Tellegen, 2008; MMPI-3; Ben-Porath & Tellegen, 2020a, b), thus lending itself to suit well to be used within a multi-method assessment battery.

Another distinctive feature of the IOP-29 is that it uses a probabilistic score to determine the level of credibility of a given presentation. The chief validity scale of the IOP-29, i.e., the

False Disorder probability Score (FDS), was indeed derived from logistic regression analyses aimed to establish the likelihood that a given IOP-29 would be obtained from a set of IOP-29s from a group of bona fide patients versus experimental feigners. The higher the FDS the lower the credibility of the IOP-29 at hand (Viglione & Giromini, 2020). Differently from the typical SVT, thus, rather than using a single set of reference data based on healthy controls, the IOP-29 uses two different sets of reference data, one coming from individuals genuinely affected by some psychiatric or cognitive disorder(s) and one coming from healthy individuals instructed to pretend to suffer from some psychiatric or cognitive disorder(s). This methodological choice aims at simplifying the decision-making process to the forensic evaluators by providing them with a direct answer to the question: is this IOP-29 more similar to those that are valid or those that are invalid? Conversely, with the typical SVT, this question is answered in a more indirect way because the interpretively relevant scores are traditionally generated by comparing the raw score of the evaluee at hand against a set of healthy controls-based reference values, and ultimately both genuinely impaired individuals and feigners/malingerers are likely to depart from these reference values.

Contributing to the cost-benefit ratio or utility of the IOP-29 is its brevity: It is typically completed within 5 to 10 minutes (Viglione & Giromini, 2020). As such, it lends itself well to any evaluation contexts in which the professional needs to work under time pressure. For instance, it can be easily added to various kinds of assessment batteries of tests or used as a brief screening measure, to decide whether additional symptom validity testing is needed. Thanks to its brevity, the IOP-29 is also suitable to be used with severely impaired individuals who would not otherwise be able to complete any long or complex instruments. Furthermore, the test can be administered both in-person and remotely, with the paper-and-pencil and computerized

administration formats generating highly comparable results (Giromini et al., 2021). Simply put, the IOP-29 can be administered in any clinical and/or forensic situations in which a negative response bias validity check is needed for clinical presentations related to neuropsychological (e.g., traumatic brain injury) and/or psychiatric (e.g., depression, anxiety, psychosis, post-traumatic stress disorder) problems. These would include, for example, psychological injury, mental state at the time of the offense, competence to stand trial, disability, immigration hardship waiver, fitness to serve time in prison, etc.

Convergent and Incremental Validity

The items of the IOP-29 were designed and refined so as to minimize their redundancy with the information provided by other existing SVTs such as the F scales of the MMPI instruments or stand-alone SVTs such as the SIMS (Viglione & Giromini, 2020). As such, in the IOP-29 manual (Viglione & Giromini, 2020), incremental validity has been prioritized over convergent validity and addressed in more detail. To overcome the lack of information concerning convergent validity and further extend that on incremental validity, we identified from our initial database comprised of 15 publications all articles in which the IOP-29 had been administered together with one or more other SVTs or performance validity tests (PVTs). If we did not have access to the original data sets, we contacted the principal investigator(s) of the relevant article(s) and asked them to share the original data set(s) with us. Then, we performed the additional convergent and incremental validity analyses presented below.

Convergent Validity. The IOP-29 includes 26 SVT-like and 3 PVT-like items. As such, the IOP-29 is *primarily* an SVT. However, as noted above, the IOP-29 is different from the typical SVT by the fact that it relies on different detection strategies. In an attempt to better understand what kind of information it delivers when included in the multi-method assessment of

symptom and performance validity, we thus reviewed the empirical literature and its pattern of correlations with other tests. We anticipated that the IOP-29 would be more highly associated with SVTs than with PVTs.

In total, we found that in 14 samples from 11 published articles the IOP-29 had been administered together with other SVTs and/or PVTs. Table 1 summarizes the convergent validity findings coming from these research efforts, obtained by re-analyzing the relevant data sets. Within the whole dataset (combined N = 2,322), the weighted mean correlation was r = .587 (weighted SD = .188).[2] This value decreased to r = .521 (weighted SD = .238) and increased to r = .653 (weighted SD = .077) when considering PVT-only or SVT-only data, respectively. Taken together, these findings suggest that the IOP-29 does correlate, with a large effect size (Cohen, 1988), with other measures or tasks that should tap the same (or a similar) constructs.

It should be noted, however, that almost all of the studies summarized in Table 1 used a simulation design so that the reported coefficients could be inflated. Indeed, simulation studies are known to artificially inflate effect sizes by the fact that in experimental contexts the feigning and control groups typically produce highly noncredible versus highly credible presentations, whereas in the actual, real-life assessments the differences between credible and noncredible profiles are typically less obvious and less bimodal (Rogers & Bender, 2018). And this is particularly true when the control group of an experimental, simulation study is comprised of healthy volunteers rather than genuinely impaired individuals, as is the case, for instance, in Gegner et al., 2021, in which the convergent validity coefficients indeed exceeded |r| = .70.

[2] To calculate the weighted mean correlation, a positive sign was assigned to correlations in the expected direction and a negative sign was assigned to correlations in the non-expected direction. For instance, the IOP-29 is supposed to correlate negatively with the TOMM and positively with the SIMS; as such, if the IOP-29 correlated r = -.50 with the TOMM and r = .30 with the SIMS, the average correlation would be r = .40.

On the other hand, it is also important to appreciate that when compared to known-groups research paradigms, simulation designs are less likely to generate a *floor effect*, in that the base rate of noncredible responding is typically higher with the latter methodological approach (Rogers & Bender, 2018; Young, 2015). Thus, when examining the results presented in Table 1 one should keep in mind not only that simulation studies may be subject to artificially inflate correlation coefficients, but also that known-groups studies are somehow more likely to artificially deflate them.

With these considerations in mind, it is worth noticing that the weakest, as well as the only nonsignificant IOP-29 correlation coefficients in Table 1, were obtained in two studies using PVT correlates. Abeare et al. (2021) used the Validity Index Seven (VI-7), a composite measure of seven embedded PVTs (Erdodi, 2019) in outpatient, comprehensive, neuropsychological evaluations. The correlate in Giromini et al. (2020a) was the Test of Memory Malingering (TOMM; Tombaugh, 1996), in a malingering experimental paradigm/simulation study without a control group. With regard to Abeare et al. (2021), one might speculate that the non-significant correlation ($r = .189$, $p = .189$) was due to its inspecting real-life evaluations, rather than using a simulation design. Such speculation, however, is inconsistent with the strong correlation between the IOP-29 and SIMS ($r = .723$, $p < .001$) observed by Roma et al. (2019) in an ecologically valid sample of court-ordered, psychological injury evaluations. A more likely explanation, thus, is that the IOP-29 simply yields a different type of information, compared to the typical PVT: As noted above, indeed, the IOP-29 is *primarily* an SVT.

Consistent with this attribution, the second of the two studies mentioned above (i.e., Giromini et al., 2020a) found that albeit the IOP-29 did not correlate with the TOMM ($-.106 \leq r \leq .017$), "both measures produced excellent sensitivity values, ranging from .82 to .98 for the

TOMM, and from .88 to 1.00 for the IOP-29 when using standard a priori cutoff scores" (p. 504). That is, even though both the IOP-29 and TOMM were sensitive to noncredible responding, they looked at symptom/performance validity from two very different angles. This explanation is also in line with emerging research indicating that non-credible scores on both SVTs and PVTs are rather uncommon, in real-life evaluations (Shura et al., 2021). Indeed, this fact reduces the importance of convergent validity as an indicator of utility for validity measures in vivo evaluations, in which incremental validity is more important.

Additionally, it is also worth pointing out that the IOP-29 breaches the most common rule of PVTs and SVTs, which is to present two alternative possible answers. The SVT-like items of the IOP-29 indeed offer three possible response option, i.e., *True*, *False*, and *Doesn't make sense*. Also, two of the three PVT-like items of the test are open-ended questions (logical and mathematical problems). By reducing the possible impact of common method variance, these differences in the response options formats thus have the potential to reduce the correlation to (and redundancy with) the typical PVTs and SVTs, too.

Incremental Validity. Two published studies aimed specifically at testing the incremental validity of the IOP-29. The first one, conducted in Italy, administered the IOP-29 and MMPI-2 to 155 adult individuals – 93 experimental feigners instructed to feign depression; 36 patients with depression in treatment; and 26 individuals assessed for possible work-related stress and deemed to be genuinely affected by depression (Giromini et al., 2019). To inspect incremental validity, the authors tested a series of logistic regression models predicting group membership (0 = control; 1 = feigner). In line with the hypothesis that the IOP-29 adds incremental validity when used in combination with the F scales of the MMPI-2, entering the IOP-29 after each of the MMPI-2 scales under consideration (i.e., F, Fb, and Fp) significantly

improved each of the tested models, $\chi^2 \geq 20.93$, $p < .001$. Furthermore, when MMPI-2 F and IOP-29 FDS z scores were summed, sensitivity increased, compared to using the MMPI-2 F scale alone, from .52 to .66 and from .38 to .60 at the specificity levels of .95 and .98, respectively. That is, using the IOP-29 in combination with the MMPI-2 F scales increased classification accuracy compared to using the MMPI-2 F scales alone.

The second of the two studies (Giromini et al., 2020a) was conducted in Portugal and investigated the extent to which the IOP-29 would yield incremental validity when used in combination with the TOMM. Using a simulation/analogue research design, the authors instructed 100 nonclinical adult volunteers to feign either depression (n = 50) or mTBI (n = 50) and examined the extent to which the IOP-29 could correctly classify as noncredible those presentations that TOMM falsely classified as credible. Confirming the hypothesis that the IOP-29 adds incremental validity when used in combination with the TOMM, all twelve false negative classifications (100%) generated by the second trial of the TOMM produced a true positive, non-credible IOP-29 result. In addition, 7 of the 8 of the false negative classifications (87%) generated by the retention trial of the TOMM obtained a true positive, non-credible IOP-29 result.

To provide more complete information on the incremental validity of the IOP-29, we retrieved the data sets from all of the studies of Table 1 that (a) used a simulation design and (b) administered at least another SVT or PVT, in addition to the IOP-29 (Table 2). For each data set, we entered the IOP-29 in the second step of a logistic regression aimed at predicting group membership (0 = control; 1 = feigner), after entering the other SVT or PVT in the first step, so as to test whether adding the IOP-29 to the model would improve the model in a statistically significant manner. Consistent with the studies described above in this section (i.e., Giromini et

al., 2019, 2020a), all incremental validity comparisons were significant, thus providing strong support for the incremental validity (i.e., improved classification accuracy) of the IOP-29 over other SVTs and various PVTs across culture, language, and type of mental health or neurological condition. More broadly, these findings confirm that the IOP-29 could be a useful addition to the toolbox of assessors performing multi-method assessment of symptom or performance validity.

Cut Scores and Hit Rates

The IOP-29 professional manual (Viglione & Giromini, 2020) suggests using FDS ≥ .50 as the standard IOP-29 cut score across psychiatric diagnoses (i.e., depression, post-traumatic stress disorder, schizophrenia/psychosis) and cognitive disorders and combinations thereof. According to the test authors, this cut score would offer sensitivity and specificity levels of about .80 across various contexts. When using the IOP-29 for screening purposes, however, professionals may consider lowering their threshold to FDS ≥ .30 to increase sensitivity, in that in such contexts only positive classifications are followed up with additional validity checks so that the risk for false negative outcomes needs to be minimized. This more liberal cut score is expected to generate a sensitivity level of approximately .90, at a specificity of approximately .60. Conversely, in high-stakes evaluations, it is a standard practice to seek specificity levels of about .90 (Sherman, Slick, & Iverson, 2020) so that the threshold for noncredible responding on the IOP-29 could be raised to FDS ≥ .65. According to Viglione and Giromini (2020), this more conservative cut score would generate a specificity of about .90 at a slightly reduced sensitivity of .70.

To evaluate the extent to which these estimates are supported by empirical research, we reviewed all published studies reporting information on IOP-29 hit rates (Table 3). Since not all of the reviewed studies presented hit rates information for all of the aforementioned cut scores,

we compensated the missing information by retrieving the data sets associated with each of the published studies and by performing additional sensitivity and specificity analyses. Additionally, Table 3 also reports Cohen's *d* effect sizes observed when comparing the average scores of the credible *versus* noncredible groups of each study. The total number of IOP-29 protocols examined in these analyses is 3,777.

Table 4 summarizes the information presented in Table 3 by calculating weighted mean hit rates values. When considering the standard IOP-29 cut score of FDS ≥ .50, the weighted mean specificity is .92 (*weighted SD* = .06; *range*: .79 – 1.00) and the weighted mean sensitivity is .86 (*weighted SD* = .07; *range*: .63 – .96). For the more liberal IOP-29 cut score of FDS ≥ .30, the weighted mean specificity is .76 (*weighted SD* = .11; *range*: .57 – .97) and the weighted mean sensitivity is .94 (*weighted SD* = .05; *range*: .70 – .99). For the more conservative IOP-29 cut score of FDS ≥ .65, the weighted mean specificity is .96 (*weighted SD* = .03; *range*: .87 – 1.00) and the weighted mean sensitivity is .76 (*weighted SD* = .08; *range*: .59 – .89). AUC values range from .83 to 1.00, with a weighted mean of .95 and a small weighted *SD* of .04. The weighted mean Cohen's *d* is 3.02 (*weighted SD* = .98; *range*: 1.48 – 5.31), an effect size value that may be characterized as *very large*, based on Rogers et al.'s (2003) suggested benchmarks.

All in all, these results compare favorably to those initially described by Viglione et al. (2017) and later summarized in the professional manual of the test (Viglione & Giromini, 2020). On the one hand, this encouraging result may be ascribed to the fact that almost all of the studies considered for these analyses were simulation studies, which – as noted above – tend to inflate effect sizes (Rogers & Bender, 2018). On the other hand, the weighted mean values reported in the previous paragraph are similar to those observed in the known-groups study by Roma et al. (2019). Indeed, the weighed mean *d* from Table 4 is 3.02; in Roma et al. (2019) *d* is 2.98.

Likewise, the weighted specificity values from Table 4 are .76, .92, and .96, respectively, for FDS ≥ .30, FDS ≥ .50, and FDS ≥ .65; in Roma et al. (2019) these values are .74, .98, and 1.00, respectively. Along similar lines, the weighted sensitivity values from Table 4 are .94, .86, and .76, respectively, for FDS ≥ .30, FDS ≥ .50, and FDS ≥ .65; in Roma et al. (2019) these values are .97, .81, and .66, respectively.

Table 4 also allows us to consider some potential moderators. More specifically, it summarizes hit rates values obtained with different study designs, different comparison groups of simulation studies, different investigated conditions, and different languages. Although the number of available studies for conducting these detailed analyses is too small to allow any conclusive statements, the small variability from one contrast to another, combined with the small weighted standard deviation values reported in Table 4 suggest that the effectiveness of the IOP-29 might generalize well from one context to another and from one type of research study to another. Although this consideration is not comprehensively and definitively evaluated, the favorable results provide initial, preliminary evidence of similar performance of the IOP-29 across diagnosis, type of study, culture, and language. Additional research on this topic, however, is sorely needed.

Strengths and Weaknesses

The IOP-29 is a brief, self-administered SVT that takes about ten minutes to complete and yet provides accurate information on the credibility of the presentation. To generate an objective yardstick for its effectiveness and precision, it would be useful to consider the results of two important meta-analytic studies that focused, respectively, on the embedded validity scales of the MMPI-2 and Personality Assessment Inventory (PAI; Morey, 1991, 2007). As for the former, Rogers et al. (2003) found that the validity of the Fp – arguably one of the most

powerful SVTs of the MMPI-2 – generated mean Cohen's *d* values of 1.90 and 2.24 in studies comparing experimental simulators against, respectively, patient or nonclinical controls. As for the latter, Hawes and Bocaccini (2009) found that the average Cohen's *d* for the PAI NIM, MAL and RDF ranged from 0.92 to 1.45 in simulation studies conducted with genuine patients as controls, and from 2.24 to 2.55 in those using nonclinical individuals as controls. In our review, the IOP-29 generated slightly higher mean Cohen's *d* values of 2.01, 2.94 and 3.59 when experimental simulators were compared, respectively, against (a) patient controls, (b) mixed clinical, forensic and/or nonclinical controls, or (c) nonclinical controls only (Table 4). Along similar lines, it would be useful to notice that one of the studies listed in Table 3 compared the effectiveness of the IOP-29 against that of the SIMS (Giromini et al., 2018), and found that the IOP-29 demonstrated significantly higher classification accuracy ($d_{IOP\text{-}29} = 1.93$ *vs.* $d_{SIMS} = 1.39$; $z = 4.57$, $p < .01$). Thus, pending additional research using criterion group designs, the IOP-29 appears to be at least as effective and possibly more effective than other established SVTs, despite its brevity.

Additionally, the IOP-29 also fits well within the multi-method assessment of symptom and/or performance validity. As demonstrated by the results reported in Table 2, indeed, it yields incremental validity both when used with other SVTs like the SIMS or the embedded validity scales of the MMPI or PAI, and when used together with a PVT like the TOMM or the FIT. One might say that the IOP-29 addresses evaluees from a different perspective, compared to other available SVTs and PVTs. As such, using it in combination with these other tools would likely help to gain a deeper understanding of the person being evaluated, compared to using the other test(s) alone.

Another major emerging strength of the IOP-29 is its adaptability and demonstration of validity across culture and language. As reviewed above, indeed, the same cut score(s) yielded similar results with different clinical and forensic populations across various cultural and assessment contexts – when considering the overall findings summarized in Table 4, the weighted standard deviation of the specificity and sensitivity values observed with the standard IOP-29 cut score of FDS ≥ .50 was ≤ .07. From a practical and applied perspective, this simplifies the use of the test and minimizes the risk of having to provide different test interpretations for the same results obtained from different evaluees (e.g., those claiming psychotic-related versus depression-related problems, etc.). Furthermore, its being suited to both online and paper-and-pencil administration formats (Giromini et al., 2021) as well as its user-friendly scoring platform (www.iop-test.com) also make the IOP-29 a very easy-to-use measure in symptom and/or performance validity assessment. All of these allow the IOP-29 to be used in the same way across many diagnoses and contexts, reducing cost in the cost-benefit ratio to increase utility.

Another strength of the IOP-29 is that it may be administered using either a paper-and-pencil or an online/computerized format. Although this aspect was not addressed directly in the current quantitative review, Giromini et al. (2021) recently inspected the comparability and validity of the online and in-person administration formats, and found that the effectiveness of the IOP-29 is preserved when alternating between face-to-face and online/remote formats. As such, it is also suitable for use in the context of tele-assessment.

Like any other test, however, the IOP-29 also has weaknesses and limitations. Unlike multi-scale, comprehensive tests, it does not measure genuine psychiatric symptoms and impairment, so that it cannot discriminate healthy from pathological bona fide test-takers. The

FDS, in other words, only informs on the level of credibility of the overall presentation; if an IOP-29 looks credible, though, additional testing with other measures is needed to evaluate the level of severity of presented symptoms/problems. For this reason, in complex forensic evaluations it would be preferable to administer the IOP-29 together with other assessment measures. More specifically, based on the data reviewed above, we speculate that it would be particularly useful to use it together with longer self-administered instruments such as the MMPI, the PAI and/or the recently introduced Self-Report of Symptom Inventory (SRSI; Merten et al., 2016). These measures, indeed, use a different approach to detecting negative response bias, compared to the IOP-29, and they all include several scales aimed at measuring genuine psychopathology.

Another limitation of the IOP-29 is that it was not designed to identify feigning of physical pain and no research has yet tested its effectiveness in a similar context. Likewise, to date, we do not know whether or not the IOP-29 would be able to identify noncredible presentations of ADHD-related problems. As such, at this time, we would not recommend using the IOP-29 in those and related settings (e.g., learning disability assessments, etc.), unless one intends to investigate their possibly related problems, such as depression or anxiety.

Furthermore, and perhaps more importantly, although the research reviewed in this article yielded quite promising results, to date almost all of the IOP-29 cross-validation studies have used a simulation design. The only exception is represented by Roma et al. (2019), who examined data from a series of court-ordered psychological evaluations. Given that, the exact extent to which the sensitivity and specificity results presented in Table 4 will really generalize to real-life medicolegal assessments is almost unknown. As such, additional IOP-29 research conducted using criterion group designs in ecologically valid contexts is highly needed, at this

time. Ideally, this future research should incorporate multiple SVT-based criteria, given that the convergent analyses discussed above clearly suggested that the IOP-29 is better characterized as an SVT, rather than as a PVT – and it is rather rare, in real-life evaluations, to find examinees who fail both types of validity checks (Shura et al., 2021). There are also not a lot of data demonstrating its validity and how it adds to PVTs in the evaluation of the credibility of cognitive disorders. Finally, additional independent research not involving the test's authors would be beneficial. Indeed, at least one of the two test authors is also a co-author of all 15 publications included in this quantitative review. Although another 46 authors from numerous different institutions from ten different countries (i.e., Australia, Brazil, Canada, England, France, Italy, Lithuania, Portugal, Slovenia, and North America) co-authored these research articles too, the IOP-29 developers likely have a better understanding of the optimal conditions of use of the instruments, as well as of how it should be presented and administered to the participant, etc. This likely has an (indirect) influence on the overall effectiveness the test has demonstrated in these studies.

Future Perspectives

Future research should also focus on two other, emerging areas of interest. A first one concerns the recently developed, IOP-29 add-on, forced-choice PVT named Inventory of Problems – Memory module (IOP-M; Giromini et al., 2020d). The IOP-M is a 34-item test that is administered immediately after the IOP-29, with the request to identify the words or brief sentences seen on the test taken immediately before (i.e., the IOP-29). Each item presents a "target" word or phrase that is in the IOP-29 questionnaire and another one that is not (a foil); for each IOP-M item, the task of the evaluee is to identify the target. Thus, the IOP-M is a forced-

choice PVT that focuses on incidental memory. It is typically completed quickly by the fact no extra time is used to present its target word and phrases.

Giromini et al.'s (2020d) had observed that (a) the IOP-29 and TOMM do not correlate with each other and (b) false negative classifications of the TOMM were correctly classified as positive cases by the IOP-29 and vice versa. With this in mind, the goal of the IOP-M is to provide valuable incremental validity when used in combination with the IOP-29. That is, using the IOP-29 together with the IOP-M is supposed to yield better classification accuracy compared to using the IOP-29 alone, especially in the evaluation of cognitive disorders. The underlying idea is that combining an SVT with a PVT likely increases signal detection overusing SVTs only or PVTs only, as demonstrated by emerging research findings (Fox & Vincent, 2020; Pivovarova et al., 2009). To date, the IOP-M effectiveness and utility, in terms of incremental validity and brevity of administration, has been demonstrated in various cultural contexts such as in Australia (Gegner et al., 2021), Brazil (Carvalho et al., 2021), France (Banovic et al., 2021), Italy (Giromini et al., 2020d), and Slovenia (Šömen et al., 2021). However, additional research is needed before confidently using it in applied clinical and forensic settings.

The second area of IOP-29 development that might deserve attention in future research concerns the recently introduced Random Responding Scale (RRS; Giromini et al., 2020c). This index aims at detecting careless, uncooperative, and/or inattentive responding and is calculated based on IOP-29 response patterns. It is designed to be independent of the FDS with the aim of identifying possible misclassifications due to content-unrelated (Nichols et al., 1989) distortions. To date, only one independent study has replicated Giromini et al.'s (2020c) findings indicating that the RRS is indeed highly elevated when test-takers are instructed to respond at random (Winters et al., 2020). Additional research on this topic would thus be beneficial.

Final Remarks

Methodological and Statistical Caveats. This quantitative review used parametric statistics to investigate the psychometric properties and effectiveness of the IOP-29. Because the FDS is a probability score, however, its values are not normally distributed so that one might wonder whether adopting a nonparametric approach would yield different results. To address this possible concern, we performed some additional analyses.

In a popular Monte Carlo simulation study by Curran, West and Finch (1996) it was observed that significant nonnormality problems are likely to arise in parametric analyses when skewness values exceed 2.0 and kurtosis values exceed 7.0. We thus inspected the skewness and kurtosis of the FDS in all credible and noncredible groups described in Table 3. Within the 16 credible groups included in this review (N = 1,826), the weighted mean skewness value was 1.44 (weighted SD = .51) and the weighted mean kurtosis value was 2.43 (weighted SD = 2.33). Within the 17 noncredible groups (N = 1,951), the weighted mean skewness value was -1.23 (weighted SD = .40) and the weighted mean kurtosis value was 1.02 (weighted SD = 2.62). As such, the distribution of scores of credible groups was slightly leptokurtic and positively skewed (i.e., most of the scores were towards the low end of the distribution), whereas that of noncredible groups was slightly leptokurtic and negatively skewed (i.e., most of the scores were towards the high end of the distribution). Based on Curran et al.'s (1996) criteria, however, these modest departures from normality should *not* generate particular problems when performing parametric analyses.

To further investigate this issue, we also tested *empirically* whether or not the application of non-parametric statistics, compared to parametric techniques, would have led to significantly different conclusions. As noted above, our convergent validity analyses based on Pearson r

revealed that when considering the whole dataset (combined *N* = 2,322), the weighted mean correlation was *r* = .587 (weighted *SD* = .188). We thus re-analyzed the same data sets also by using the non-parametric analog, i.e., Spearman *Rho*. When looking at the results of these new, non-parametric analyses, the results were remarkably similar, nearly identical, with a weighted mean rank-order correlation of *Rho* = .598 (weighted *SD* = .187). Along similar lines, the parametric statistic used to quantify the extent to which the IOP-29 scores from the credible versus noncredible groups differ from each other is Cohen's *d*. The *AUC* is an analog non-parametric effect size estimator (Ruscio and Mullen, 2012). In this quantitative review, both indexes (weighted mean *d* = 3.02; weighted mean *AUC* = .95) concurred to indicate that the IOP-29 has excellent validity in discriminating credible from noncredible symptom presentations. Overall, it can therefore be concluded that the use of parametric or non-parametric techniques leads to completely comparable results.

Test Security and Interpretive Caveats. The detection strategies utilized by the IOP-29 are introduced and generally described in the original test developmental article (Viglione et al., 2017) and in the IOP-29 professional manual (Viglione & Giromini, 2020). Neither source, however, reproduces the exact details of the mathematical formulas that are needed to generate the FDS by hand. This is because the divulgation of the scoring algorithm used by the official scoring program of the IOP-29 to produce the FDS (www.iop-test.com) could potentially jeopardize the overall utility and effectiveness of the test. After all, given that the test is so brief, should the scoring algorithm be released to the broader audience, it would likely become overly easy to coach clients on how to feign mental illness on the IOP-29 without being identified as feigners.

Another issue that deserves mention is that the IOP-29 is currently available in 11 languages, and several other linguistic and cultural adaptations are being developed. Each translation/adaptation has been developed in collaboration with the test authors, by following a standard translation/back-translation procedure (Brislin, 1980; Geisinger, 2003; Van de Vijver & Hambleton, 1996). As such, we recommend professionals to only use these official translations (which can be downloaded from the IOP-29 website at www.iop-test.com) and avoid interpreting "on-the-spot" generated translations.

Conclusions. The current review article represents the first systematic effort to summarize all of the published research investigating psychometric characteristics and validity of the IOP-29. Taken together, the data reviewed support the recent characterization of the IOP-29 as "a newer stand-alone SVT that has the required psychometric properties for use in forensic disability and related assessments. Its research profile is accumulating, a hallmark for use in legal settings" (Young et al.'s 2020, p. 460)."

References

References marked with an asterisk indicate records included in the quantitative review.

* Abeare, K., Razvi, P., Sirianni, C.D., Giromini, L., Holcomb, M., Cutler, L., Kuzmenka, P., & Erdodi, L.A. (2021). Introducing Alternative Validity Cutoffs to Improve the Detection of Non-credible Symptom Report on the BRIEF. *Psychological Injury and Law, 14*, 2–16. https://doi.org/10.1007/s12207-021-09402-4

*Ales, F., Giromini, L., Warmelink, L., Polden, M., Wilcockson, T., Kelly, C., Winters, C., Zennaro, A., & Crawford, T. (2021). An Eye Tracking Study on Feigned Schizophrenia. *Psychological Injury and Law, 14(3),* 213-226

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.

* Banovic, I., Filippi, F., Viglione, D.J., Scrima, F., Zennaro, A., Zappalà, A., & Giromini, L. (2021). Detecting Coached Feigning of Schizophrenia with the Inventory of Problems – 29 (IOP-29) and Its Memory Module (IOP-M): A Simulation Study on a French Community Sample. *International Journal of Forensic Mental Health, [Epub ahead of print]*. https://doi.org/10.1080/14999013.2021.1906798

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

Ben-Porath, Y. S., & Tellegen, A. (2020a). *MMPI-3 Manual for Administration, Scoring, and Interpretation*. Minneapolis, MN: University of Minnesota Press.

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

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

* Carvalho, L., Reis, A., Colombarolli, M.S., Pasian, S.R., Miguel, F.K., Erdodi, L.A., Viglione, D.J., & Giromini, L. (2021). Discriminating Feigned from Credible PTSD Symptoms: a Validation of a Brazilian Version of the Inventory of Problems-29 (IOP-29). *Psychological Injury and Law, 14*, 58-70. https://doi.org/10.1007/s12207-021-09403-3

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

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

Fox, K.A., & Vincent, J.P. (2020). Types of Malingering in PTSD: Evidence from a Psychological Injury Paradigm. *Psychological Injury and Law 13*, 90–104 (2020). https://doi.org/10.1007/s12207-019-09367-5

* Gegner, J., Erdodi, L.A., Giromini, L., Viglione, D.J., Bosi, J. & Brusadelli, E. (2021). An Australian study on feigned mTBI using the Inventory of Problems – 29 (IOP-29), its Memory Module (IOP-M), and the Rey Fifteen Item Test (FIT). *Applied Neuropsychology: Adult, [Epub ahead of Print]*. https://doi.org/10.1080/23279095.2020.1864375

* Giromini, L., Barbosa, F., Coga, G., Azeredo, A., Viglione, D. J., & Zennaro, A. (2020a). Using the inventory of problems - 29 (IOP-29) with the test of memory malingering (TOMM) in symptom validity assessment: A study with a portuguese sample of experimental feigners. *Applied Neuropsychology: Adult, 27*, 504-516. https://doi.org/10.1080/23279095.2019.1570929

* Giromini, L., Carfora Lettieri, S. C., Zizolfi, S., Zizolfi, D., Viglione, D. J., Brusadelli, E., Zennaro, A. (2019). Beyond rare-symptoms endorsement: A clinical comparison simulation study using the minnesota multiphasic personality inventory-2 (MMPI-2) with the inventory of problems-29 (IOP-29). *Psychological Injury and Law, 12*, 212-224. https://doi.org/10.1007/s12207-019-09357-7

Giromini, L., Pignolo, C., Young, G., Drogin, E.Y., Zennaro, A., & Viglione, D.J. (2021). Comparability and Validity of the Online and In-Person Administrations of the Inventory of Problems-29. *Psychological Injury and Law, [Epub ahead of print]*. https://doi.org/10.1007/s12207-021-09406-0

* Giromini, L., Viglione, D. J., Pignolo, C., & Zennaro, A. (2018). A clinical comparison, simulation study testing the validity of SIMS and IOP-29 with an Italian sample. *Psychological Injury and Law*, *11,* 340-350. https://doi.org/10.1007/s12207-018-9314-1

* Giromini, L., Viglione, D. J., Pignolo, C., & Zennaro, A. (2020b). An Inventory of Problems - 29 sensitivity study investigating feigning of four different symptom presentations via malingering experimental paradigm. *Journal of Personality Assessment, 102*, 563-572. https://doi.org/10.1080/00223891.2019.1566914

Giromini, L., Viglione, D. J., Pignolo, C., & Zennaro, A. (2020c). An Inventory of Problems - 29 (IOP-29) study on random responding using experimental feigners, honest controls, and

computer-generated data. *Journal of Personality Assessment, 102*, 731-742. https://doi.org/10.1080/00223891.2019.1639188

* Giromini, L., Viglione, D. J., Zennaro, A., Maffei, A., & Erdodi, L. A. (2020d). SVT Meets PVT: Development and Initial Validation of the Inventory of Problems – Memory (IOP-M). *Psychological Injury and Law, 13*, 261–274. https://doi.org/10.1007/s12207-020-09385-8

Hawes, S.W., & Boccaccini, M.T. (2009). Detection of Overreporting of Psychopathology on the Personality Assessment Inventory: A Meta-Analytic Review. *Psychological Assessment, 21*(1), 112-124. https://doi.org/10.1037/a0015036

* Ilgunaite, G., Giromini, L., Bosi, J., Viglione, D. J., & Zennaro, A. (2020). A clinical comparison simulation study using the Inventory of Problems-29 (IOP-29) with the Center for Epidemiologic Studies Depression Scale (CES-D) in Lithuania. *Applied Neuropsychology: Adult, [Epub ahead of print]*. https://doi.org/10.1080/23279095.2020.1725518

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.

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*, 102–111.

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

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

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology, 45(2),* 239–250. https://doi.org/10.1002/1097-4679(198903)45:2<239::AID-JCLP2270450210>3.0.CO;2-1

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.

Pivovarova, E., Rosenfeld, B., Dole, T., Green, D. & Zapf, P. (2009). Are Measures of Cognitive Effort and Motivation Useful in Differentiating Feigned from Genuine Psychiatric Symptoms?. *International Journal of Forensic Mental Health, 8(4),* 271-278. https://doi.org/10.1080/14999011003635514

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. (2018). An introduction to response styles. In R. Rogers & Bender, S.D. (Eds.), *Clinical assessment of malingering and deception* (4th ed). (pp. 3–17). New York, NY: Guilford Press.

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

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

* Roma, P., Giromini, L., Burla, F., Ferracuti, S., Viglione, D. J., & Mazza, C. (2019). Ecological validity of the Inventory of Problems-29 (IOP-29): an Italian study of court-ordered, psychological injury evaluations using the Structured Inventory of Malingered Symptomatology (SIMS) as criterion variable. *Psychological Injury and Law, 13*, 57-65. https://doi.org/10.1007/s12207-019-09368-4

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research, 47(2)*, 201–223. https://doi.org/10.1080/00273171.2012.658329

Sherman, E. M. S., Slick, D. J., & Iverson, G. L. (2020). Multidimensional malingering criteria for neuropsychological assessment: A 20-year update of the malingered neuropsychological dysfunction criteria. *Archives of Clinical Neuropsychology, 35*(6), 735-764. https://doi.org/10.1093/arclin/acaa019

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

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.

* Šömen, M.M., Lesjak, S., Majaron, T., Lavopa, L., Giromini, L., Viglione, D.J., & Podlesek, A. (2021). Using the Inventory of Problems-29 (IOP-29) with the Inventory of Problems Memory (IOP-M) in Malingering-Related Assessments: a Study with a Slovenian Sample of Experimental Feigners. *Psychological Injury and Law, [Epub ahead of print]*. https://doi.org/10.1007/s12207-021-09412-2

Tombaugh, T. N. (1996). *Test of Memory Malingering (TOMM)*. New York, USA: Multi Health Systems.

* 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

Viglione, D. J., Giromini, L., Landis, P., McCullaugh, J.M., Pizitz, T.D., O'Brien, S., Wood, S., Connell, K., & Abramsky, A. (2019). Development and Validation of the False Disorder Score: The Focal Scale of the Inventory of Problems. *Journal of Personality Assessment, 101*, 653-661. https://doi.org/10.1080/00223891.2018.1492413

Viglione, D.J., & Giromini, L. (2020). *Inventory of Problems–29: Professional Manual*. Columbus, OH: IOP-Test, LLC.

* Winters, C. L., Giromini, L., Crawford, T. J., Ales, F., Viglione, D. J., & Warmelink, L. (2020). An Inventory of Problems–29 (IOP–29) study investigating feigned schizophrenia and random responding in a British community sample. *Psychiatry, Psychology and Law, [Epub ahead of print]*. https://doi.org/10.1080/13218719.2020.1767720

Wood, S. (2008). *Unique contributions of performance and self-report methods in the detection of malingered psychotic symptoms* (Unpublished doctoral dissertation). California School of Professional Psychology, San Diego, CA.

Young, G. (2015). Malingering in forensic disability-related assessments: Prevalence 15±15%. *Psychological Injury and Law*, *8*(3), 188-199. https://doi.org/10.1007/s12207-015-9232-4.

Young, G., Foote, W.E., Kerig, P.K., Mailis, A., Brovko, J., Kohutis, E.A., McCall, S., Hapidou, E.G., Fokas, K.F., Goodman-Delahunty, J. (2020). Introducing Psychological Injury and Law. *Psychological Injury and Law, 13*, 452–463. https://doi.org/10.1007/s12207-020-09396-5

Table 1. Convergent Validity of the IOP-29.

Source	Sample	*N*	Other Measure(s)	Type of Measure	*r*	*p*
Abeare et al. (2021)	Adult individuals from Canada, referred for neuropsychological evaluations	50	VI-7	PVT	.189	.189
Abramsky' subsample of Viglione et al. (2017)	Experimental feigners of depression & patients w/depression from the US	85	TOMM-1	PVT	-.380	< .001
			TOMM-2	PVT	-.452	< .001
Banovic et al. (2021)	Experimental feigners of schizophrenia & community-based controls from France	114	IOP-M	PVT	-.637	< .001
Carvalho et al. (2021)	Experimental feigners of PTSD & community-based controls from Brazil	201	IOP-M	PVT	-.433	< .001
Gegner et al. (2021)	Experimental feigners of mTBI & community-based controls from Australia	275	FIT	PVT	-.717	< .001
			IOP-M	PVT	-.736	< .001
Giromini et al. (2018)	Experimental feigners of various conditions & patients w/various diagnoses from Italy	452	SIMS	SVT	.689	< .001
Giromini et al. (2019)	Experimental feigners of depression & patients w/depression from Italy	155	MMPI-2 F	SVT	.654	< .001
			MMPI-2 Fb	SVT	.695	< .001
			MMPI-2 Fp	SVT	.469	< .001
Giromini et al. (2020a)	Experimental feigners of depression or mTBI from Portugal	100	TOMM-1	PVT	.017	.870
			TOMM-2	PVT	-.096	.342
			TOMM-r	PVT	-.106	.294

Giromini et al. (2020d)	Experimental feigners of various conditions & community-based controls from Italy	360	IOP-M	PVT	-.672	< .001
McCullaugh' subsample of Viglione et al. (2017)	Offenders on probation from the US, instructed to feign mental illness or respond honestly	128	PAI NIM	SVT	.720	< .001
			PAI MAL	SVT	.547	< .001
			PAI RDF	SVT	.610	< .001
O'Brien' subsample of Viglione et al. (2017)	Experimental feigners of psychosis & patients w/psychosis from the US	88	MMPI-2 F	SVT	.636	< .001
			MMPI-2 Fp	SVT	.601	< .001
			MMPI-2 Ds-r2	SVT	.735	< .001
Roma et al. (2019)	Adult individuals from Italy, tested for possible psychological injury (court-ordered evaluations)	74	SIMS	SVT	.723	< .001
Šömen et al. (2021)	Experimental feigners of schizophrenia or depression & community-based controls from Slovenia	150	IOP-M	PVT	-.583	< .001
Wood' subsample of Viglione et al. (2017)	Experimental feigners of psychosis & patients w/psychosis from the US	90	PAI NIM	SVT	.746	< .001
			PAI MAL	SVT	.729	< .001
			PAI RDF	SVT	.591	< .001

Notes. VI-7 = Validity Index Seven – a composite measure consisting of seven embedded PVTs (Erdodi, 2019); TOMM-1 = Test of Memory Malingering, Trial 1; TOMM-2 = Test of Memory Malingering, Trial 2; TOMM-r = Test of Memory Malingering, Retention Trial; IOP-M = Inventory of Problems – Memory module; FIT = Fifteen Item Test; SIMS = Structured Inventory of Malingered Symptoms; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; PAI = Personality Assessment Inventory.

Table 2. Incremental Validity of the IOP-29.

Source	Honest Controls		Experimental Feigners		SVT/PVT Entered at Step 1	χ^2 of the Model at Step 1	χ^2 of the Model at Step 2 (w/IOP-29)	$\Delta\chi^2$
	Characterization	*n*	Characterization	*n*				
Abramsky' subsample of Viglione et al. (2017)	Patients w/depression from the US	43	Experimental feigners of depression from the US	42	TOMM-1	26.8	67.1	40.3**
					TOMM-2	40.2	71.8	31.5**
Gegner et al. (2021)	Community-based controls from Australia	93	Experimental feigners of mTBI from Australia	182	FIT	180.9	331.6	150.6**
Giromini et al. (2018)	Patients w/various diagnoses from Italy	216	Experimental feigners of various conditions from Italy	236	SIMS	170.2	269.8	99.6**
Giromini et al. (2019)[a]	Patients w/depression from Italy	62	Experimental feigners of depression from Italy	93	MMPI-2 F	81.1	105.1	24.0**
					MMPI-2 Fb	72.2	93.1	20.9**
					MMPI-2 Fp	45.4	94.5	49.1**
McCullaugh' subsample of Viglione et al. (2017)	Offenders on community-based probation from the US (controls)	64	Offenders on probation instructed to feign various conditions from the US (feigners)	64	PAI NIM	120.8	146.8	26.0**
					PAI MAL	50.5	121.6	71.1**
					PAI RDF	72.4	126.2	53.8**
O'Brien' subsample of Viglione et al. (2017)	Patients w/psychosis from the US	43	Experimental feigners of psychosis from the US	45	MMPI-2 F	19.5	37.0	17.5**
					MMPI-2 Fp	25.0	40.1	15.2**
					MMPI-2 Ds-r2	45.8	49.8	4.0*

Wood' subsample of Viglione et al. (2017)	Patients w/psychosis from the US	45	Experimental feigners of psychosis from the US	45	PAI NIM	46.2	60.3	14.0**
					PAI MAL	54.4	68.6	14.2**
					PAI RDF	50.1	73.3	23.2**

Notes. TOMM-1 = Test of Memory Malingering, Trial 1; TOMM-2 = Test of Memory Malingering, Trial 2; FIT = Fifteen Item Test; SIMS = Structured Inventory of Malingered Symptoms; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; PAI = Personality Assessment Inventory. All models were statistically significant both at step 1 and at step 2 at $p < .01$. * $p < .05$; ** $p < .01$. [a] Results from this study have been published before in Giromini et al. (2019).

Table 3. Hit Rates of the IOP-29: Individual Studies.

Source	Credible Group		Noncredible Group		FDS ≥ .30		FDS ≥ .50		FDS ≥ .65		*AUC*	Co
	Characterization	***n***	**Characterization**	***n***	**Sp**	**Se**	**Sp**	**Se**	**Sp**	**Se**		
Abeare et al. (2021)	Students from Canada	46	Experimental feigners of cognitive impairment from Canada	27	.93	.70	.98	.63	1.00	.59	.83	1
Ales et al. (2021)	Community sample from England	40	Experimental feigners of schizophrenia from England	43	.90	.93	.98	.88	1.00	.86	.98	3
Banovic et al. (2021)	Community sample from France	37	Experimental feigners of schizophrenia from France	77	.81	.86	.92	.73	.97	.60	.89	1
Carvalho et al. (2021)	Firefighters exposed to potentially traumatic event(s) (*n* = 154) & community sample (*n* = 101) from Brazil	255	Experimental feigners of PTSD from Brazil	100	.68	.97	.89	.87	.97	.69	.95	2
Gegner et al. (2021)	Community sample from Australia	93	Experimental feigners of mTBI from Australia	182	.94	.98	1.00	.96	1.00	.89	1.00	5
Giromini et al. (2018)	Patients w/various diagnoses from Italy	216	Experimental feigners of various disorders from Italy	236	.60	.90	.82	.81	.93	.73	.89	1

Giromini et al. (2019)	Credible evaluees (*n* = 26) & patients w/depression (*n* = 36) from Italy	62	Experimental feigners of depression from Italy	93	.71	.89	.87	.75	.89	.67	.89	1
Giromini et al. (2020a)	N/A	0	Experimental feigners of depression (*n* = 50) or mTBI (*n* = 50) from Portugal	100	N/A	.97	N/A	.92	N/A	.82	N/A	1
Giromini et al. (2020b)[a]	Community sample from Italy	400	Experimental feigners of various disorders from Italy	400	.76	.96	.93	.91	.97	.83	.96	3
Giromini et al. (2020d)	Community sample from Italy, w/elderly responders (*n* = 32) likely suffering from cognitive impairment	192	Experimental feigners of various disorders from Italy	168	.82	.97	.94	.86	.99	.72	.98	3
Ilgunaite et al. (2020)	Patients w/depression from Lithuania	50	Experimental feigners of depression from Lithuania	50	.72	.98	.96	.94	.98	.74	.98	3
McCullaugh' subsample of Viglione et al. (2017)	Offenders on community-based probation from the US (controls)	64	Offenders on probation instructed to feign various conditions from the US (feigners)	64	.97	.80	1.00	.72	1.00	.66	.94	2
Roma et al. (2019)	Credible forensic evaluees (SIMS score < 17) from Italy	43	Noncredible forensic evaluees (SIMS score ≥ 17) from Italy	32	.74	.97	.98	.81	1.00	.66	.98	2

Šömen et al. (2021)	Community sample from Slovenia	50	Experimental feigners of depression (*n* = 50) or schizophrenia (*n* = 50) from Slovenia	100	.88	.97	.98	.88	.98	.73	.99
Viglione et al.'s (2017) cross-validation sample	Patients w/various diagnoses from the US	82	Experimental feigners of various disorders from the US	83	.57	.93	.79	.81	.90	.69	.87
Winters et al. (2020)[a]	Community sample from England	151	Experimental feigners of schizophrenia from England	151	.89	.99	.97	.92	.97	.83	.99
Wood' subsample of Viglione et al. (2017)	Patients w/psychosis from the US	45	Experimental feigners of schizophrenia from the US	45	.67	.96	.80	.82	.87	.69	.90

Notes. Sp = Specificity; Se = Sensitivity. SIMS = Structured Inventory of Malingered Symptoms. [a] This study used a within-subject design, in which participants were asked to take the IOP-29 three times, one time answering honestly, one time faking mental illness, and one time responding with a random-like approach.

Table 4. Hit Rates of the IOP-29: Weighted Mean Values.

	Specificity					Sensitivity					Effect size			
	N	***k***	**FDS ≥ .30**	**FDS ≥ .50**	**FDS ≥ .65**	***N***	***k***	**FDS ≥ .30**	**FDS ≥ .50**	**FDS ≥ .65**	***N***	***k***	***AUC***	***d***
Overall	1,826	16	.76 *(.11)*	.92 *(.06)*	.96 *(.03)*	1,951	17	.94 *(.05)*	.86 *(.07)*	.76 *(.08)*	3,677	16	.95 *(.04)*	3.02 *(0.99)*
Study Design														
Simulation	1,783	15	.76 *(.11)*	.91 *(.06)*	.96 *(.03)*	1,919	16	.94 *(.05)*	.86 *(.07)*	.76 *(.08)*	3,602	15	.95 *(.04)*	3.02 *(1.00)*
Criterion Group	43	1	.74	.98	1.00	32	1	.97	.81	.66	75	1	.98	2.98
Comparison Group of Simulation Studies														
Nonclinical Sample	817	7	.83 *(.07)*	.95 *(.03)*	.98 *(.01)*	980	7	.95 *(.05)*	.89 *(.07)*	.81 *(.08)*	1,797	7	.96 *(.04)*	3.59 *(0.98)*
Clinical Sample	455	5	.63 (.05)	.83 (.05)	.92 (.03)	507	5	.92 *(.03)*	.81 *(.05)*	.71 *(.03)*	962	5	.90 *(.03)*	2.01 *(0.45)*
Mixed/Other	511	3	.77 *(.10)*	.92 *(.04)*	.98 *(.01)*	332	3	.94 *(.07)*	.84 *(.06)*	.70 *(.02)*	843	3	.96 *(.02)*	2.94 *(0.28)*

Target Condition of the Study

Depression/Anxiety	112	2	.71 *(.00)*	.91 *(.04)*	.93 *(.04)*	143	2	.92 *(.04)*	.82 *(.09)*	.69 *(.03)*	255	2	.93 *(.04)*	2.39 *(0.74)*
PTSD	255	1	.68	.89	.97	100	1	.97	.87	.69	355	1	.95	2.71
Psychosis	273	4	.84 *(.08)*	.94 *(.06)*	.96 *(.04)*	316	4	.95 *(.05)*	.85 *(.08)*	.76 *(.10)*	589	4	.96 *(.04)*	3.34 *(1.08)*
Neuropsychological	139	2	.94 *(.00)*	.99 *(.01)*	1.00 *(.00)*	209	2	.94 *(.09)*	.92 *(.11)*	.85 *(.10)*	348	2	.96 *(.07)*	4.51 *(1.56)*
Mixed/Other	1,047	7	.74 *(.11)*	.91 *(.06)*	.96 *(.03)*	1,183	8	.94 *(.04)*	.86 *(.05)*	.76 *(.06)*	2,130	7	.94 *(.04)*	2.81 *(0.63)*
Language														
English	521	7	.84 *(.14)*	.94 *(.08)*	.96 *(.05)*	595	7	.94 *(.08)*	.87 *(.09)*	.79 *(.10)*	1,116	7	.95 *(.06)*	3.54 *(1.41)*
Italian	913	5	.73 *(.08)*	.90 *(.05)*	.96 *(.03)*	929	5	.94 *(.03)*	.86 *(.05)*	.76 *(.06)*	1,842	5	.94 *(.04)*	2.79 *(0.63)*
Other	392	4	.72 *(.07)*	.91 *(.03)*	.97 *(.00)*	427	5	.95 *(.04)*	.87 *(.07)*	.72 *(.07)*	719	4	.95 *(.03)*	2.80 *(0.52)*

Notes. Values in parenthesis represent the weighted SDs.

Figure 1. PRISMA Four-Phase Search Strategy Flow Diagram.