

Psychometric Review of the Personality Style Test

Sabina Alispahić, Ph.D.

Professor of Clinical and Health Psychology
Department of Psychology, Faculty of Philosophy
University of Sarajevo
ORCID: 0009-0005-6740-6106

I conducted a psychometric review of the IDRLabs [Personality Style Test](#) using the supplied results package covering 105 items, 15 proposed personality-style scales, and all 481,730 de-duplicated test sittings logged between 30 July and 27 October 2026. The purpose was to evaluate data quality, score distribution, internal consistency, item behaviour, and the extent to which the 15 proposed styles function as coherent and distinguishable scales.

481,730	105	15	.91
Analysed sittings	Items	Proposed styles	Overall alpha

1. Data quality

The original export contained 484,726 logged records. After same-visitor identical-answer duplicates were collapsed, 481,730 de-duplicated sittings remained, and all of them could be scored. No sample was drawn: every de-duplicated sitting enters the analyses below. The test was represented across 19 published locales, led by English (41%), Russian (25%) and Spanish (17%).

Each item is a statement answered Agree or Disagree, and each style is scored from seven items as the percentage answered Agree. Within the analysed data, 1,363 sittings (0.28%) gave the same answer to every item. This is a small proportion and is unlikely to determine the aggregate findings. However, no time-based or other response-quality exclusion rule was applied. The median completion time is 11 minutes 46 seconds, approximately 6.7 seconds per item in the 319,347 timed sittings, which is compatible with reading each statement; future validation should nevertheless pre-specify response-quality criteria and test whether the results are robust to reasonable exclusions.

The unit of analysis is the sitting rather than the person: duplicates were removed only when the same visitor submitted identical answers, so a person who retook the test with different answers contributes more than once. The package supplied for this review contains aggregate psychometric outputs, item statistics, and generated figures rather than the session-level response matrix. The statistics below therefore describe and interpret the supplied outputs; raw-data sensitivity analyses cannot be independently recomputed from this package alone.

The sample is very large for an initial evaluation, but participants are self-selected internet users. These data should not be treated as representative population norms, estimates of the prevalence of any personality disorder, or evidence of diagnostic accuracy.

2. Score distribution

The test reports fifteen style scores rather than a total, so the overall score examined here (the percentage of all 105 items answered Agree) is an analytic summary rather than a result shown to respondents. It is close to symmetric: mean = 54.3/100, median = 55.2, SD = 15.2, skewness = -0.27, and excess kurtosis = -0.05. The 5th and 95th percentiles are 28 and 77, and the interquartile range runs from 44 to 66.

This is a well-behaved distribution with substantial differentiation among respondents and no floor or ceiling compression. Because the overall score counts agreement across all fifteen styles, it is best read as a general index of endorsed personality-style features and, since every item is scored in the Agree direction, partly as an index of the tendency to agree (Section 5). Given the self-selected sample, it should not be interpreted as a population distribution of personality pathology.

The fifteen style means range from 41.5 (Histrionic) to 68.9 (Depressive), with standard deviations of 23.9 to 32.6, so every style spreads respondents widely. Eleven of the fifteen style means are at or above 50. The result page names a single style when the highest style score is at least 50% (four of seven items) and is not tied; since scores at that level are common in this sample, the threshold should be understood as a descriptive convention rather than an empirically derived cut-off.

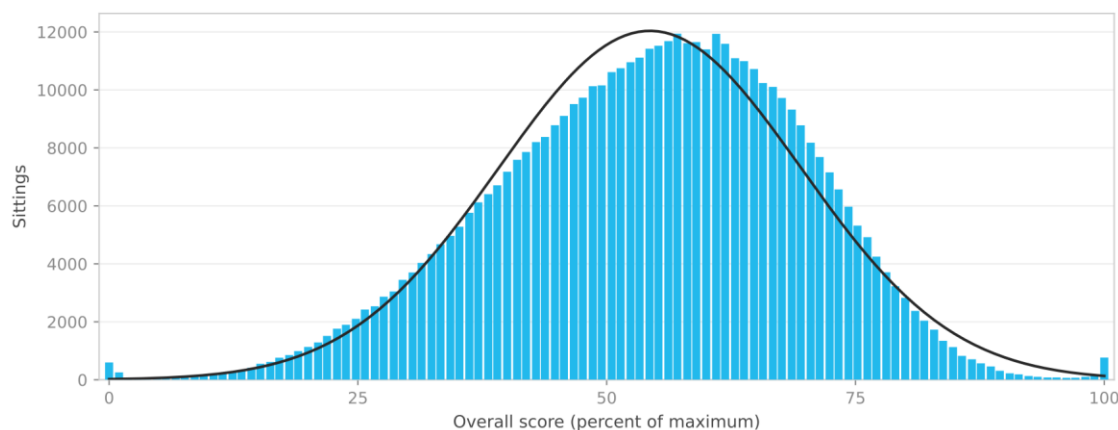


Figure 1. Overall-score distribution (percentage of the 105 items answered Agree) across all 481,730 analysed sittings, from the supplied results package.

3. Internal consistency and item behaviour

Internal consistency for the full 105-item set is high: Cronbach's alpha (KR-20, as the items are binary) = .91. With 105 items a high value is expected even when the average inter-item correlation is modest, and because the items span fifteen partly opposed styles, it supports a broad common dimension in the item pool rather than showing that the total measures a single clinical construct.

Style reliability ranges from alpha = .52 to .79 (median .68). Six of the fifteen seven-item scales reach .70 or above and six fall between .63 and .69; three are lower: Negativistic

(.56), Compulsive (.56) and Hypomaniac (.52). For seven agree/disagree items these values are not unusual, but the three lowest scales are not yet reliable enough for confident individual-level interpretation and justify item-level review.

Style	α	Style	α
Narcissistic	.68	Depressive	.76
Antisocial	.63	Masochistic	.79
Histrionic	.76	Borderline	.72
Dependent	.72	Paranoid	.69
Sadistic	.68	Schizoid	.66
Compulsive	.56	Schizotypal	.67
Negativistic	.56	Hypomaniac	.52
Avoidant	.72		

Every item correlates positively with the rest of its own style: corrected item-style correlations range from $r = .18$ to $.60$ (median $.38$), so there is no obvious item behaving in the wrong direction. Agreement rates range from 23% to 85%, so no item sits at the floor or ceiling. Twenty-one items fall below $r = .30$ with their own style and three below $.20$: two Hypomaniac items (extravagant clothing or make-up, $r = .18$; an urge to drive fast, $r = .19$) and one Negativistic item (anger followed by guilt, $r = .19$). These, together with the weaker items of the three lowest-reliability scales, would be reasonable candidates for wording review or replacement testing.

Against the overall score, corrected item-total correlations range from $r = -.09$ to $.52$. The seven negative values belong to three Histrionic and four Compulsive items. Because each of these items correlates positively with its own style, this reflects how those two styles relate to the others (Section 4) rather than a scoring error.

4. Style structure

The fifteen style scores are unevenly related, with pairwise correlations ranging from $r = -.33$ to $.67$ (median $.26$); 21 of the 105 pairs are negative. Avoidant, Depressive, Masochistic, Borderline and Schizotypal form a tight cluster ($r = .53$ to $.67$), and Dependent, Negativistic and Paranoid correlate with that cluster at $r = .39$ to $.54$. Narcissistic, Antisocial and Sadistic form a second, smaller cluster ($r = .51$ to $.53$). Histrionic correlates negatively with Avoidant ($r = -.33$), Schizoid ($-.31$), and Masochistic and Depressive (both about $-.20$), and Compulsive is nearly independent of every other style ($|r| \leq .17$).

The eigenvalues of the style correlation matrix make the same point: the first two components account for 35% and 17% of the style-score variance (eigenvalues 5.3 and 2.6), two further components exceed 1 (1.4 and 1.1), and the remaining eleven are each .65 or below. This pattern is compatible with a few broad dimensions plus some style-specific information. On the present evidence, a small number of broad dimensions is better supported than a claim that all fifteen styles are distinct latent constructs; at the same time,

the weak or negative relations of Histrionic and Compulsive with the rest mean that a single total does not summarise the instrument well either.



Figure 2. Style-score distributions, means, standard deviations, and internal-consistency estimates from the supplied results package.

Factor analysis of the item-level data, with estimators suited to binary responses, is needed to compare a fifteen-factor representation with models built on a few broad dimensions, and with higher-order and bifactor models. Such work would clarify which styles add reliable information beyond the broad dimensions.

5. Construct interpretation and validation priorities

At the content level, the instrument samples fifteen personality styles drawn from personality-style theory, including the work of Theodore Millon. Ten correspond by name to the ten personality disorders of DSM-5-TR (Compulsive to obsessive-compulsive personality disorder); the other five (Sadistic, Masochistic, Negativistic, Depressive and Hypomaniac) are not current diagnostic categories, and ICD-11 has replaced categorical personality-disorder types with a severity rating and trait domains. The available results support the internal coherence of a self-report inventory of these tendencies.

Style scores are not diagnoses of personality disorders. A personality-disorder diagnosis requires an enduring, pervasive pattern with clinically significant distress or impairment,

established through clinical assessment; seven agree/disagree statements per style cannot establish that, and the current package does not show that any style score, or the 50% result threshold, has known sensitivity or specificity for a diagnosis. Internal consistency and item discrimination are necessary psychometric properties, but they are not substitutes for external or clinical validation.

Three characteristics of the instrument and sample deserve particular attention. 86% of the 321,363 device-identified sittings were completed on mobile, and the mobile mean overall score is 1.9 points higher than the desktop mean (54.7 vs 52.7). The difference is small, and device type may be confounded with language, age, or country, but invariance across devices should still be tested. Third, English, Russian and Spanish together account for 83% of sittings across 19 locales, so translation equivalence and cross-language measurement invariance are important before pooled multilingual norms are interpreted strongly.

Further validation should ideally examine:

- convergent validity against established measures of personality pathology (for example, the PID-5 and Millon-type inventories) and structured clinical interviews such as the SCID-5-PD;
- criterion validity against structured diagnostic assessment, including sensitivity, specificity, and empirically derived cut-offs in place of the fixed 50% threshold if the test is to be used as a screener;
- discriminant validity against current depression and anxiety symptoms, general distress, and normal-range personality traits;
- the contribution of acquiescence, using reverse-keyed items or acquiescence-controlled models;
- factor analysis of item-level data comparing fifteen-factor, broad-dimension, higher-order, and bifactor models;
- test-retest reliability;
- measurement invariance across language, sex, age, device type, and other relevant groups; and
- pre-specified response-quality and completion-speed sensitivity analyses in an independent sample.

6. Overall conclusion

The initial internal psychometric evidence for the Personality Disorder Test is encouraging, but it remains preliminary. Across 481,730 sittings, the 105-item instrument shows high overall internal consistency (.91), acceptable reliability for twelve of its fifteen seven-item style scales (.63 to .79), positive discrimination for every item within its own style, agreement rates well away from floor and ceiling, and well-spread score distributions.

The main unresolved issues concern structure, response style, and external and clinical validity rather than obvious basic scoring failures. Most styles overlap within one large cluster while two stand largely apart, three scales (Hypomaniac, Compulsive and Negativistic) have modest reliability.

Subject to those limitations, the available evidence supports continued development and formal validation of the measure as a self-report inventory of personality-style tendencies. It would be premature, on the supplied evidence alone, to treat any style score as a diagnosis of a personality disorder, a population norm, or an estimate of disorder prevalence.

The Personality Style Test thus has promising internal psychometrics from a very large sample, with the strongest support presently attaching to a few broad dimensions rather than Millon's fifteen separate styles. Style scores describe self-reported tendencies, not personality-disorder diagnoses; structural, acquiescence-controlled, external, and clinical validation should come next.

Signature:

A handwritten signature in blue ink, appearing to read "J. M. Millon" with a stylized flourish at the end.

Date: 28.10.2026.