Psychometric Review of the Body Dysmorphic 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

14 September 2026

I conducted a preliminary psychometric review of the IDRLabs Body Dysmorphic Test using the supplied results package covering 32 items, eight proposed facets, and an analysed random sample of 10,000 test sittings drawn from 18,196 de-duplicated sessions. The purpose was to evaluate data quality, score distribution, internal consistency, item behaviour, and the extent to which the eight proposed facets function as coherent and distinguishable components of the broader measure.

10,000	32	8	.95
ANALYSED SITTINGS	ITEMS	PROPOSED FACETS	OVERALL ALPHA

1. Data quality

The original export contained 18,315 logged records. After same-visitor identical-answer duplicates were collapsed, 18,196 de-duplicated sessions remained. A fixed random sample of 10,000 sittings was used for the detailed item and facet analyses, and the test was represented across 16 published locales.

Within the analysed sample, 17 sittings (0.17%) used the same response option for 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 reported median completion time is 3 minutes 8 seconds, approximately 5.9 seconds per item in the timed subset, so future validation should pre-specify response-quality criteria and test whether the results are robust to reasonable exclusions.

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 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 body dysmorphic disorder, or evidence of diagnostic accuracy.

2. Score distribution

The total score shows broad dispersion but a clear concentration toward the upper part of the scale: mean = 70.6/100, median = 73, SD = 17.9, skewness = -1.00, and excess kurtosis = 1.18. The 5th and 95th percentiles are 35 and 94, respectively.

This is not a degenerate distribution: scores span a substantial portion of the scale and there is meaningful differentiation among respondents. At the same time, the negative skew and high upper percentiles indicate appreciable high-end compression. Given the self-selected nature of the sample, this may reflect who chooses to take the test as much as it reflects the instrument itself, and it should not be interpreted as a population symptom distribution.

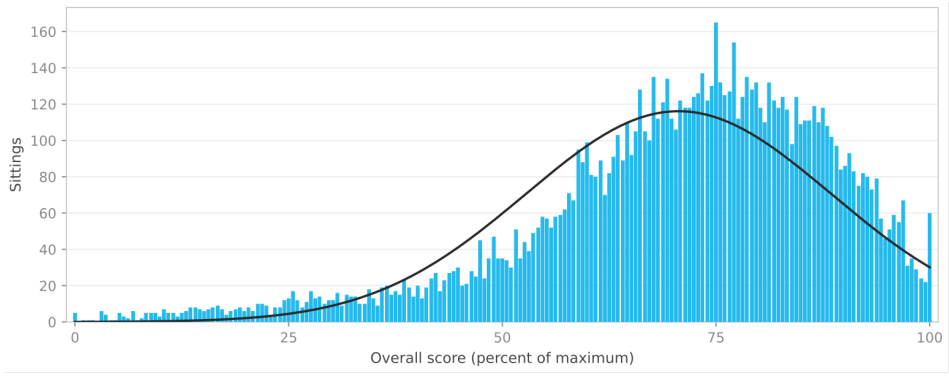

Figure 1. Overall-score distribution in the analysed 10,000-sitting sample from the supplied results package.

3. Internal consistency and item behaviour

Internal consistency is very strong for the full 32-item measure: Cronbach's alpha = .95. This supports the coherence of the total score, although an alpha this high can also reflect overlap or redundancy among items and should not be mistaken for evidence of clinical validity.

Facet reliability ranges from alpha = .65 to .84. Seven of the eight four-item facets fall between .71 and .84; Excessive Self-Scrutiny is lower at .65. Because every facet contains only four items, the lower value should not be interpreted mechanically, but it does justify item-level review and replication.

Facet	α	Facet	α
Preoccupation with Appearance	**.83**	Avoidance of Social Situations	**.84**
Excessive Self-Scrutiny	**.65**	Impaired Functioning	**.78**
Excessive Comparisons	**.73**	Negative Quality of Life	**.84**
Repetitive Behaviors	**.71**	Warped Self-image	**.79**

Corrected item-total correlations range from r = .35 to .76, and the supplied first-component loadings range from .36 to .80 after reverse-keying. All are positive in the scored direction, so there is no obvious item that is behaving in the wrong direction. A small number of items, particularly within Excessive Self-Scrutiny and Repetitive Behaviors, show weaker discrimination near r = .35-.40 and would be reasonable candidates for wording review or replacement testing.

4. Facet structure

The eight facet scores are moderately to strongly intercorrelated, with pairwise correlations ranging from r = .42 to .84. Several relationships are high, including Preoccupation with Appearance with Negative Quality of Life (r = .84) and Impaired Functioning with Negative Quality of Life (r = .81).

This pattern is compatible with a strong broader body-dysmorphic symptom dimension, while also leaving room for some domain-specific information. However, the correlations are high enough that the distinct value of reporting eight separate subscales should be tested rather than assumed. On the present evidence, the total score is psychometrically better supported than a claim that all eight facets are sharply distinct latent constructs.

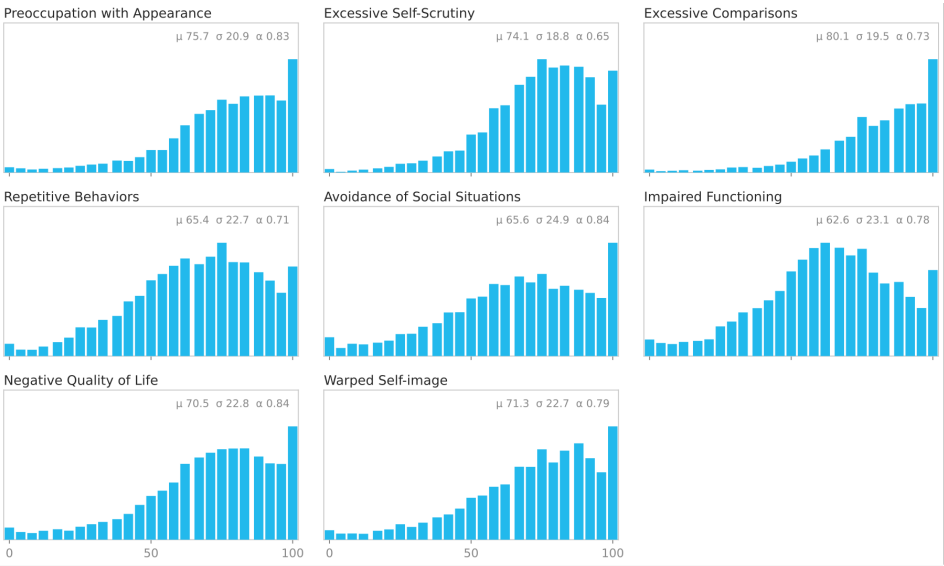

Figure 2. Facet-score distributions, means, standard deviations, and internal-consistency estimates from the supplied results package.

Confirmatory factor analysis using item-level data is needed to compare an eight-factor representation with simpler one-factor, higher-order, and bifactor models. Such work would clarify whether the facets add reliable information beyond a common general factor.

5. Construct interpretation and validation priorities

At the content level, the instrument samples several themes that are clinically relevant to body-dysmorphic symptom reporting: appearance preoccupation, checking and self-scrutiny, comparison, repetitive appearance-related behaviours, social avoidance, impaired functioning, reduced quality of life, and distorted self-image. The available results support the internal coherence of a self-report measure of these experiences.

The current package does not establish that the score is a diagnosis of body dysmorphic disorder, that any cut-off has known sensitivity or specificity, or that the scale distinguishes body dysmorphic disorder from related forms of distress. Internal consistency and item discrimination are necessary psychometric properties, but they are not substitutes for external or clinical validation.

Two sample characteristics deserve particular attention. First, 95% of the device-identified sittings were completed on mobile, and the reported mobile mean is 5.7 points higher than the desktop mean (71.1 vs 65.5). This does not by itself demonstrate a mode effect because device type may be confounded with language, age, country, or other factors, but the difference is large enough to justify formal invariance testing. Second, the locale distribution is uneven, with Russian accounting for 76% of sittings in the package. Translation equivalence and cross-language measurement invariance are therefore important before pooled multilingual norms are interpreted strongly.

Further validation should ideally examine:

- convergent validity against established body-dysmorphic symptom measures and clinician-rated assessment;
- criterion validity against structured diagnostic evaluation, including sensitivity, specificity, and empirically derived cut-offs if the test is to be used as a screener;
- discriminant validity against general body dissatisfaction, eating pathology, depression and anxiety, obsessive-compulsive symptoms, and social anxiety;
- confirmatory factor analysis comparing one-factor, eight-factor, higher-order, and bifactor models;
- test-retest reliability and sensitivity to change;
- 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 Body Dysmorphic Test is encouraging, but it remains preliminary. The 32-item instrument demonstrates very strong overall internal consistency, generally acceptable-to-good reliability across most four-item facets, positive item discrimination throughout the supplied item set, and a total-score distribution with substantial range.

The main unresolved issues concern structural, external, and clinical validity rather than obvious basic scoring failures. The eight facets overlap substantially, the score distribution is concentrated toward the high end, and the observed device and language composition make subgroup and mode effects important to investigate. Several lower-discriminating items also merit targeted review.

Subject to those limitations, the available evidence supports continued development and formal validation of the measure as a self-report screening instrument. It would be premature, on the supplied evidence alone, to treat the score as a clinical diagnosis, a population norm, or a validated estimate of disorder prevalence.

> **Bottom line:** promising internal psychometrics for a large-sample first pass, with the strongest support presently attaching to the overall score. External, clinical, and structural validation should come next.

Signature:

Sabina Alispahić, Ph.D.

Date: 14.9.2026.