

Psychometric Review of the Autism Spectrum Test

Jennifer Schulz, Ph.D., MSW, LSW

BSW Program Director | Associate Professor
Ashland University

21 August 2026

I conducted a preliminary psychometric review of the IDRlabs Autism Spectrum Test using datasets comprising 50 items and 135,823 valid test sessions collected across the languages in which the test is offered. The purpose was to evaluate data quality, score distribution, internal consistency, item behavior, and the functioning of the ten proposed facets.

1. Data quality

The dataset is very large for an initial psychometric evaluation. The original dataset contained 136,171 records. After duplicate sittings were identified and removed, 135,823 valid sessions remained. No sittings were excluded for incomplete or malformed responses.

The available data include individual item responses, ten facet scores, overall scores, test language, and limited geographic and technical information where available. Test-takers are identified only through a pseudonymous visitor identifier, with no names or contact information contained in the dataset.

A total of 381 participants, approximately 0.28% of the sample, selected the same response option for all 50 items. These records were deliberately retained. Because such response patterns may represent careless, protest, or otherwise non-substantive responding, future analyses should ideally specify an a priori response-quality criterion.

Importantly, excluding these straight-line responders has virtually no effect on the principal reliability estimate.

The sample size provides a strong basis for psychometric analysis. However, it is a self-selected internet sample and should not be interpreted as a representative normative sample of either the general population or individuals with clinically diagnosed autism.

2. Score distribution

The overall score distribution is statistically well behaved:

- Mean = 50.88/100
- Median = 51
- SD = 17.25
- Skewness = 0.02

- Excess kurtosis = -0.51

The overall distribution is therefore essentially symmetrical, although somewhat flatter than a normal distribution.

Scores span almost the complete available scale, from 2 to 100, without substantial overall floor or ceiling compression.

Several isolated peaks in the score distribution correspond precisely to straight-line response patterns. The principal concentrations occur at scores of approximately 8, 50, and 92. Because straight-line responders constitute only 0.28% of the dataset, these peaks do not materially determine the broader distribution.

The mean facet scores range from approximately 41 for Abnormal/Flat Speech to 57 for Anxiety, indicating meaningful variation between the proposed domains rather than a uniform response profile across all facets.

3. Internal consistency

Internal consistency for the complete 50-item instrument is high:

Cronbach's $\alpha = .937$

After removal of the 381 straight-line responders:

Cronbach's $\alpha = .936$

The negligible difference indicates that the strong overall reliability is not being artificially produced by straight-line response patterns.

Reliability of the individual five-item facets is more variable:

- Depression: $\alpha = .410$
- Fixations: $\alpha = .687$
- Abnormal/Flat Speech: $\alpha = .610$
- Noise Sensitivity: $\alpha = .808$
- Social Difficulty: $\alpha = .622$
- Anxiety: $\alpha = .639$
- Anormal Posture: $\alpha = .665$
- Poor Eye Contact: $\alpha = .867$
- Tics and Fidgets: $\alpha = .797$
- Aggression: $\alpha = .767$

Several facets therefore demonstrate good internal consistency for very brief five-item measures, particularly Poor Eye Contact, Noise Sensitivity, Tics and Fidgets, and Aggression.

Other facets show moderate reliability and may benefit from item refinement.

The Depression facet requires particular attention.

4. Item behavior

Most items show positive associations with the overall scale. Corrected item-total correlations extend to approximately $r = .61$, with many items falling in a useful moderate range.

However, several items warrant review.

Most notably, Item 5 has a corrected item-total correlation of approximately $r = -.37$ after the supplied reverse scoring. This is an unusual result for an item intended to belong to the same construct as the remainder of the test.

The effect is also apparent within the Depression facet. With the supplied scoring, Depression has an internal consistency of only $\alpha = .410$. When Item 5 is omitted, reliability of the remaining four Depression items rises to approximately $\alpha = .722$.

This suggests that Item 5 should be examined closely for its wording, intended direction, scoring key, or conceptual fit with the Depression facet before further validation.

Items 30 and 50 also display relatively weak corrected item-total relationships, approximately $r = .13$ and $r = .07$, respectively. Item 12 is somewhat stronger but remains comparatively weak at approximately $r = .23$.

These findings do not invalidate the instrument, but they provide clear empirical targets for item-level revision.

5. Relationship between the facets

The ten facet scores are positively and moderately intercorrelated.

Across facet pairs, correlations range approximately from:

$r = .26$ to $r = .64$

with an average correlation of approximately:

$r = .47$

This pattern is broadly compatible with a multidimensional instrument in which the proposed domains are related but not interchangeable.

The substantial common variance among facets also helps explain the high internal consistency of the total score.

However, internal consistency and facet correlations cannot establish the proposed dimensional structure by themselves. A factor-analytic investigation is therefore an important next step.

In particular, confirmatory factor analysis should test whether the data are best represented by ten correlated facets, a broader general factor, a hierarchical model, or another structure.

6. Interpretation and clinical limitations

The present findings demonstrate several useful psychometric properties.

The overall score has a broad and well-behaved distribution, the complete instrument shows high internal consistency, several of the individual facets show good reliability, and most items contribute positively to the broader scale.

At the same time, these data concern the internal statistical properties of the instrument. They do not establish that the test can diagnose autism.

The dataset does not contain independent clinical diagnoses or scores from established autism measures. It therefore cannot establish diagnostic sensitivity, specificity, predictive value, clinically meaningful cutoff scores, or agreement with clinician-administered assessment.

This distinction is particularly important for a self-report measure addressing autism-related characteristics.

Further validation should therefore include:

- confirmatory factor analysis of the proposed ten-facet structure;
- revision or investigation of Item 5 and other low-discrimination items;
- convergent validity against established autism measures;
- discriminant validity against related psychological constructs;
- comparison with independently assessed clinical and non-clinical groups;
- test-retest reliability;
- examination of measurement invariance across test languages and relevant populations.

7. Overall conclusion

The initial psychometric evidence for the Autism Spectrum Test is encouraging at the level of the overall instrument.

The 50-item measure demonstrates high overall internal consistency, a broad and approximately symmetrical score distribution, and substantial differentiation among respondents. Several of its five-item facets also demonstrate good internal reliability.

The analysis additionally identifies specific areas where refinement is warranted. Most importantly, Item 5 behaves inconsistently with its intended scoring and substantially reduces the reliability of the Depression facet. Several other items show comparatively weak discrimination and should also be examined.

The very large dataset provides an unusually strong foundation for further psychometric development, particularly factor analysis and item refinement.

At the present stage, the evidence supports describing the instrument conservatively as a self-report measure of autism-related traits and experiences, rather than as a diagnostic instrument.

Establishing clinical or diagnostic interpretation will require external validation against established measures and independently assessed clinical samples.

Signature:

A handwritten signature in black ink, appearing to read 'Jennifer Schulz', written over the printed name.

Jennifer Schulz, Ph.D., MSW, LSW

21 August 2026

