

Psychometric Review of the IDRLabs Jealousy Test

Dr. Jon Gruda

Invited Associate Professor of Organisational Behaviour

Católica Porto SBE, Porto, Portugal, and Maynooth University, Ireland

ORCID: 0000-0002-8315-0826

13 October 2026

Source basis: IDRLabs Jealousy Test results package and respondent-level response file (367,818 cleaned sittings; 24 items; 3 facets), analysed in Stata 19, the live test page, and the published Multidimensional Jealousy Scale.

I reviewed the IDRLabs Jealousy Test using the results package and the respondent-level response file IDRLabs supplied. Together they cover 24 items, three facets (Cognitive, Emotional and Behavioral Jealousy) and 367,818 cleaned sittings recorded between 30 July and 9 September 2026. I also compared the item wording and response format of the live test with the Multidimensional Jealousy Scale (MJS) of Pfeiffer and Wong (1989), on which the test is based.

I re-estimated every statistic from the response file in Stata 19, and the recomputed facet and overall scores match the live engine's stored scores for all 367,818 sittings. The recomputed means, standard deviations, alphas, item statistics and facet correlations agree with the results package to within 0.1 point, and the small differences arise because the package averages the engine's whole-number facet scores.

1. Data quality

The raw log held 369,429 result-page renders. Collapsing reloads by the same visitor with identical answers removed 1,611 of them, leaving 367,818 sittings, all of which the live engine could grade. The file holds no names or contact details, and visitors are identified only by a salted pseudonymous identifier.

Careless responding is rare and does not affect the results. Of the cleaned sittings, 7,060 (1.92%) gave the same answer to all 24 items, and they produce the spikes at 100 ($n = 5,313$), 0 ($n = 1,188$) and 50 ($n = 347$) in Figure 1. Among the 240,565 timed sittings, the median completion time was 2 minutes 49 seconds (7.0 seconds per item). Only 1,718 (0.71%) took less than 2 seconds per item, and 1,005 of these were also straight-line sittings. Excluding both groups (360,045 sittings remain) moves the overall mean from 57.9 to 57.4, the SD from 22.7 to 22.1 and overall alpha from 0.943 to 0.939. Facet alphas change by at most 0.013, and no item statistic changes direction. The factor analyses in sections 4 and 5 exclude straight-line sittings.

The sample is large but self-selected and unevenly spread across languages and devices. Russian-language sittings make up 67% of the total, Spanish 21% and English 8%, and 96% of sittings with a device signal came from mobile phones. Mean overall scores differ by language (Russian 61.6, English 52.2, Spanish 48.9) and by device (mobile 60.9, desktop 52.2). The pooled norms describe people who take this test on this site and should not be read as population norms for any country or language.

2. Score distribution

Overall scores spread across nearly the full 0 to 100 range.

- Mean = 57.9/100
- Median = 59
- SD = 22.7
- Skewness = -0.257
- Excess kurtosis = -0.615
- 5th and 95th percentiles = 18 and 93

The distribution is mildly negatively skewed and flatter than a normal curve. At the overall level there is almost no floor or ceiling crowding. Only 0.3% of sittings score 0 and 1.4% score 100, and every one of them is a straight-line sitting, because only a uniform response pattern can produce either extreme.

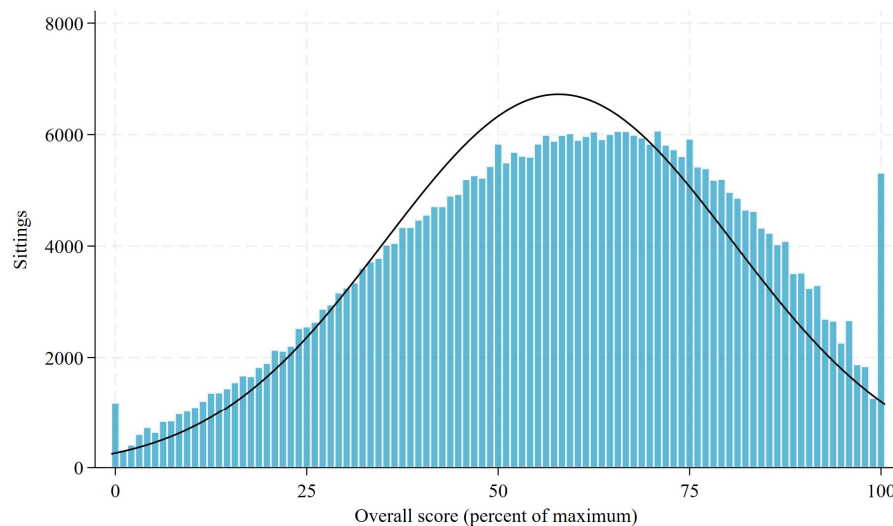


Figure 1. Overall-score distribution, all 367,818 sittings. Bars count sittings at each possible score. The curve is a normal distribution with the same mean and standard deviation.

The facets sit at different levels, with means of 58.8 for Cognitive, 69.7 for Emotional and 45.2 for Behavioral Jealousy. Takers endorse feeling upset at a partner's attention to others more readily than suspicion, and suspicion more readily than checking or monitoring. The Emotional facet has a ceiling effect, with 10.8% of sittings at the maximum score of 100 and 26.9% at 90 or above (skewness -0.77). Excluding straight-line sittings still leaves 9.5% at the maximum, against 4.7% for Cognitive and 0.8% for Behavioral Jealousy. The Emotional facet therefore separates people poorly at its upper end.

3. Reliability and item behaviour

Internal consistency is high for the full scale and for each facet, and scores are reasonably stable over time (Table 1).

Scale	Items	Cronbach's α	SEM (points)	95% band	Retest r
Full scale	24	0.943	5.4	± 10.6	0.77
Cognitive Jealousy	8	0.895	8.8	± 17.3	0.75
Emotional Jealousy	8	0.891	8.2	± 16.0	0.74
Behavioral Jealousy	8	0.857	9.3	± 18.3	0.71

Table 1. Alpha on all 367,818 sittings. $SEM = SD \times \sqrt{1 - \alpha}$, in points on the 0 to 100 score scale, and the 95% band is $\pm 1.96 \times SEM$. Retest r is the correlation between first and second sittings at least 7 days apart ($n = 881$, median interval 10.1 days).

For IDRLabs the standard errors matter more than alpha, because the site returns an individual score to every taker. A single facet score carries a 95% band of about ± 16 to ± 18 points. Two facet scores for the same person need to differ by about 24 points before the difference is unlikely to reflect measurement error alone, and IDRLabs should not present smaller differences between a taker's facets as meaningful.

The response file identifies 15,185 respondents who took the test more than once. Most of them (11,580) retook it within an hour, and agreement between those sittings is lower ($r = 0.55$ for the overall score) than between sittings a day or more apart ($r = 0.71$, $n = 2,370$) or a week or more apart ($r = 0.77$, $n = 881$). People who retake the test immediately probably change their answers on purpose, for example to answer about a different relationship, so the longer intervals give the better estimate. Mean scores rose by only 1.4 points between sittings a week or more apart. All of these figures are lower bounds, because the cleaning step merged repeat sittings with identical answers.

All corrected item-total correlations are positive (0.39 to 0.70), as are all standardised loadings in the three-factor model (0.42 to 0.79). Item 19 is the weakest item on every index ($r = 0.39$, loading 0.42), and its wording departs from the MJS. The original item asks about a partner's "previous or present romantic relationships". The IDRLabs version drops "present", so it now asks about curiosity over a partner's past, which differs from monitoring a current relationship. Item 8, the lowest Cognitive item ($r = 0.58$), was also rewritten, from suspecting that the partner "is crazy about members of the opposite sex" to thinking that the partner "cannot control their physical urges to have sex with someone else". Item 17 (looking through

a partner's drawers, bag or pockets) has the lowest mean of any item, 0.93 on the 0 to 4 scale, and discriminates mainly among high scorers, which is acceptable for an item that describes the most intrusive form of checking.

4. Factor structure

I compared four confirmatory models on the 360,758 sittings without straight-line responding, treating the five-point items as continuous (Table 2). At this sample size chi-square rejects every model, so the comparison rests on the approximate fit indices.

Model	CFI	TLI	RMSEA	SRMR
One factor	0.79	0.77	0.102	0.068
Three correlated factors (current scoring)	0.89	0.87	0.075	0.052
Four correlated factors (Cognitive split)	0.93	0.92	0.060	0.050
Bifactor (general + three facets)	0.95	0.94	0.052	0.045

Table 2. Maximum-likelihood confirmatory factor analyses, $N = 360,758$. The four-factor model splits the Cognitive facet into partner suspicion (items 1, 3, 4, 7, 8) and worry about rivals (items 2, 5, 6). The bifactor model has a general factor and three orthogonal facet factors.

A single factor fits poorly, and so does the three-factor model that matches how IDRLabs scores the test (CFI 0.89, RMSEA 0.075). The Cognitive items fall into two clusters, suspicion that the partner is involved with someone else (items 1, 3, 4, 7 and 8) and worry that other people are pursuing the partner (items 2, 5 and 6). In the bifactor model below, the two clusters load on the Cognitive factor with opposite signs, with item 3 near zero. Splitting the Cognitive facet in two raises CFI from 0.89 to 0.93. This split matches the short form of the MJS developed by Elphinston, Feeney and Noller (2011), which keeps the five partner-suspicion items and drops the three rival items.

The bifactor model fits best (CFI 0.95, RMSEA 0.052, SRMR 0.045). The general factor accounts for 70% of the common variance (explained common variance, ECV = 0.70) and omega hierarchical for the total score is 0.85, which makes the overall score a strong measure of a single general jealousy dimension. Once the general factor is removed, the facets retain much less reliable variance of their own. The Cognitive facet retains almost none (omega hierarchical for the subscale = 0.020), the Emotional facet a modest amount (0.265) and the Behavioral facet the most (0.401). In the three-factor model the latent facet correlations are 0.80 (Cognitive and Emotional), 0.72 (Cognitive and Behavioral) and 0.75 (Emotional and Behavioral). In practice the Cognitive facet score mostly repeats the overall score, and IDRLabs should not describe a high or low Cognitive score as a distinct finding about the taker.

The shared response format may contribute to the strength of the general factor. In the MJS of Pfeiffer and Wong (1989), each facet has its own 7-point scale. Respondents rate how often they have the Cognitive thoughts and Behavioral actions, and rate their reaction to each Emotional situation from “very pleased” to “very upset”. The IDRLabs version puts all 24 items under the stem “When I am in a relationship, I often...” with one 5-point scale from Disagree to Agree. A common stem and scale add shared method variance. The format also mixes frequency with agreement, since a person who does something occasionally has no obvious answer to “I often...”.

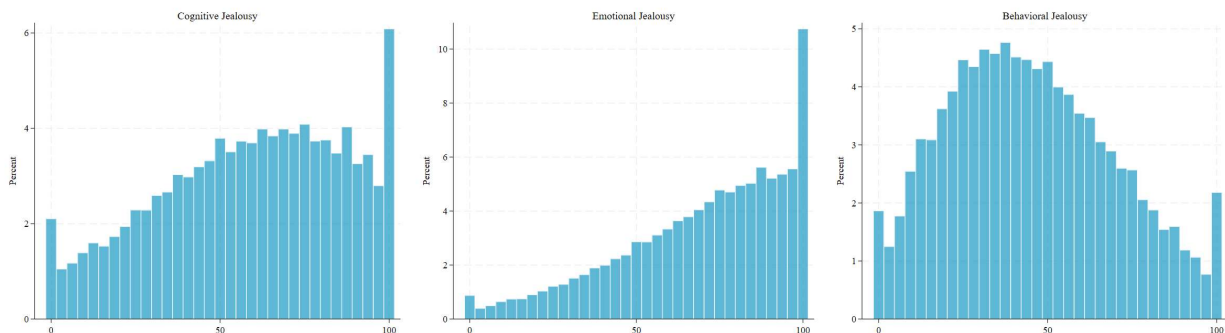


Figure 2. Facet-score distributions, all 367,818 sittings, as the percentage of sittings at each possible score.

5. Measurement invariance across device and language

I tested invariance with multi-group confirmatory factor analysis of the three-factor scoring model, comparing a configural model with a metric model (equal loadings) and a scalar model (equal loadings

and intercepts). Following Chen (2007), I treat a drop in CFI of more than 0.010 as evidence against invariance (Table 3). The same tests on the four-factor model lead to the same conclusions.

Groups and model	CFI	RMSEA	SRMR	Δ CFI
Language, configural	0.8894	0.0728	0.0549	–
Language, metric	0.8867	0.0717	0.0623	–0.0027
Language, scalar	0.8455	0.0816	0.0665	–0.0412
Device, configural	0.8982	0.0709	0.0510	–
Device, metric	0.8982	0.0695	0.0529	0.0000
Device, scalar	0.8977	0.0683	0.0529	–0.0005

Table 3. Multi-group maximum-likelihood confirmatory factor analyses of the three-factor model, excluding straight-line sittings. Language groups are Russian ($n = 240,641$), Spanish ($n = 75,944$) and English ($n = 28,851$). Device groups are mobile ($n = 225,845$) and desktop ($n = 10,577$). Δ CFI is the change from the preceding model.

The test is fully invariant across devices, so mobile and desktop scores can be compared directly. CFI stays at 0.898 as loadings and then intercepts are constrained to be equal, and no item behaves differently on the two devices (largest matched difference 0.15 item SD). Desktop takers score lower on all three latent facets ($d = -0.34$ to -0.40), and adjusting for language and country reduces the raw 8.7-point gap to 5.0 points, so part of the gap reflects who uses which device.

Across the Russian, Spanish and English versions, metric invariance holds (Δ CFI = -0.003), so the items relate to the facets in the same way in each language. Scalar invariance does not hold (Δ CFI = -0.041), because some items are easier to endorse in one language than in another at the same level of jealousy. Stata's score tests single out item 14 (being bothered when someone asks the partner out on a date). Releasing its intercept constraint gives a χ^2 of about 126,000 on 2 df, against 33,651 for the next item, item 7. The mean of item 14 is 2.13 in Spanish against 3.43 in Russian and 3.07 in English, which points to a translation problem. Matched on the rest of the scale, Spanish takers differ from Russian takers by at least 0.25 item SD on nine items, led by item 14 at -0.83 SD. English and Russian takers differ on three items, by at most 0.31 SD. These differences largely cancel in the total. They shift the overall score by less than one point (-0.72 for Spanish and -0.54 for English takers), so the 12.7-point gap between Russian and Spanish takers does not come from item bias. Facet scores and single items are not yet equivalent across languages, and IDRLabs should not compare them across languages or publish language-specific norms until the Spanish translation has been reviewed.

6. Content and construct interpretation

The IDRLabs test adapts the MJS rather than reproducing it. Besides the response format, the adaptation changes the frame of reference. The MJS asks respondents about one named current partner, whereas the IDRLabs stem asks about relationships in general, so the test measures a tendency across relationships. That choice suits a public self-test, because many takers are not in a relationship. IDRLabs should say so on the result pages. The adaptation also rewords items 8, 13, 14 and 19 and replaces the MJS surprise-visit item with item 24 (spying on a partner's email or chats).

Seven items (9, 10, 11, 12, 15, 16 and 23) keep the 1989 wording "the opposite sex", which does not fit takers in same-sex relationships. Replacing it with "someone else" or "another person" would remove the problem without changing what the items ask.

I recommend the following further work.

- Report the overall score as the primary result. Either present the Cognitive facet as part of the general score or split it into partner suspicion and worry about rivals. A separate rival scale would need more than its current three items.
- Restore the original wording of item 19 ("previous or present"), replace "the opposite sex" in seven items, and add one or two harder Emotional items to reduce the ceiling on that facet.
- Review the Spanish translation, starting with item 14 and then items 2, 3, 7, 18 and 22, and repeat the invariance test before publishing language-specific norms.
- Test convergent validity against the original MJS or its short form, and discriminant validity against attachment anxiety, general suspiciousness and relationship satisfaction.
- Collect a scheduled retest, for example two weeks after the first sitting, to replace the opportunistic retest estimates above.

Interpretive boundary

The evidence supports descriptive use of the overall score for self-reflection, and of the Emotional and Behavioral facets as secondary information. It does not support population norms, clinical interpretation, comparisons of facet scores across languages, or the claim that a taker's facets differ unless the difference exceeds about 24 points.

7. Overall conclusion

The IDRLabs Jealousy Test is a reliable measure of a general jealousy dimension. Overall alpha is 0.94, omega hierarchical is 0.85, the retest correlation over a week or more is 0.77, and the results hold when careless and very fast sittings are removed. Every item loads positively on the general factor, and full scalar invariance holds across mobile and desktop.

The facet scores carry less distinct information than the scoring implies. The three-factor structure that the scoring assumes fits poorly, and the Cognitive facet adds almost nothing beyond the general factor while splitting into partner and rival clusters. The Emotional facet has a ceiling, with roughly one sitting in ten at the maximum score. Scalar invariance does not hold across languages, and item 14 in the Spanish version shows the largest difference. The remaining problems concern wording, since item 19 has lost part of its original meaning and seven items assume a partner of the opposite sex. None of these requires a change to the overall score, and most can be fixed through wording, translation review and how facet results are presented.

Overall assessment

The overall score is reliable ($\alpha = 0.94$, retest $r = 0.77$) and works the same way on mobile and desktop, which makes it suitable for descriptive self-assessment. IDRLabs should present the facet scores more cautiously, the Cognitive facet in particular. Priority fixes are the wording of item 19, gender-neutral wording for seven items, the Spanish translation of item 14 and harder Emotional items.

References

- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Elphinston, R. A., Feeney, J. A., & Noller, P. (2011). Measuring romantic jealousy: Validation of the Multidimensional Jealousy Scale in Australian samples. *Australian Journal of Psychology*, 63(4), 243–251. <https://doi.org/10.1111/j.1742-9536.2011.00026.x>
- Pfeiffer, S. M., & Wong, P. T. P. (1989). Multidimensional jealousy. *Journal of Social and Personal Relationships*, 6(2), 181–196. <https://doi.org/10.1177/026540758900600203>

Signature: 

Dr. Jon Gruda

Invited Associate Professor of Organisational Behaviour
Católica Porto SBE, Porto, Portugal, and Maynooth University, Ireland
ORCID: 0000-0002-8315-0826

13 October 2026

Prepared from the IDRLabs Jealousy Test results package and response file. This is a psychometric review of the test. It is not a clinical diagnosis or a statement of population norms.