

01 / CONTENT BLUEPRINT

Shorter surveys. Visible tradeoffs.

A reproducible six-page audit using public-domain IPIP item wording and simulated responses. Calibration N = 800; holdout N = 400; seed = 2026.

This demonstration combines the 10-item IPIP Anxiety, Anger and Depression facets. It is a custom 30-item demonstration bank, **not** an established IPIP emotional-distress scale. The six-item candidate removes 80% of questions; clinical adequacy is not established.

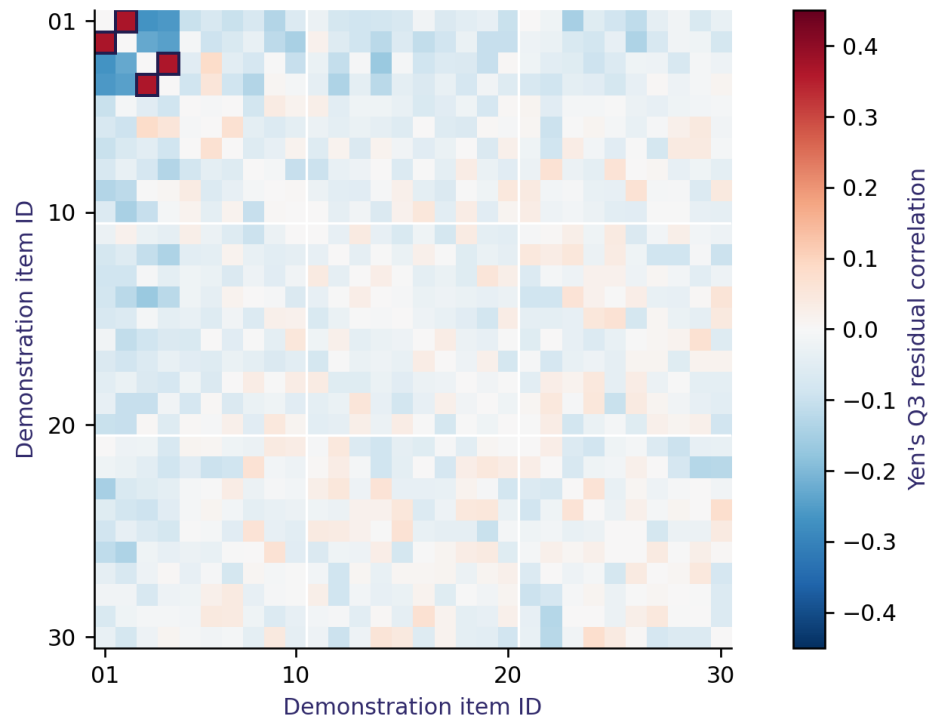
Demo ID	IPIP item wording	Facet / key	Six?
01	Worry about things.	Anxiety / +	-
02	Fear for the worst.	Anxiety / +	Yes
03	Am afraid of many things.	Anxiety / +	Yes
04	Get stressed out easily.	Anxiety / +	-
05	Get caught up in my problems.	Anxiety / +	-
06	Am not easily bothered by things.	Anxiety / -	-
07	Am relaxed most of the time.	Anxiety / -	-
08	Am not easily disturbed by events.	Anxiety / -	-
09	Don't worry about things that have already happened.	Anxiety / -	-
10	Adapt easily to new situations.	Anxiety / -	-
11	Get angry easily.	Anger / +	-
12	Get irritated easily.	Anger / +	-
13	Get upset easily.	Anger / +	-
14	Am often in a bad mood.	Anger / +	Yes
15	Lose my temper.	Anger / +	-
16	Rarely get irritated.	Anger / -	-
17	Seldom get mad.	Anger / -	Yes
18	Am not easily annoyed.	Anger / -	-
19	Keep my cool.	Anger / -	-
20	Rarely complain.	Anger / -	-
21	Often feel blue.	Depression / +	-
22	Dislike myself.	Depression / +	Yes
23	Am often down in the dumps.	Depression / +	-
24	Have a low opinion of myself.	Depression / +	-
25	Have frequent mood swings.	Depression / +	-
26	Feel desperate.	Depression / +	-
27	Feel that my life lacks direction.	Depression / +	-
28	Seldom feel blue.	Depression / -	-
29	Feel comfortable with myself.	Depression / -	-
30	Am very pleased with myself.	Depression / -	Yes

Keys reproduce the original IPIP direction (+ / -). For real 0-4 response data, reverse-keyed items require 4 - response. Here, simulated categories are **already keyed** toward the trait: do not reverse them again. DEMO IDs are local identifiers, not official IPIP item numbers.

02 / REDUNDANCY AND CONTENT

When consistency conceals overlap

Yen's Q3 residual correlations from the training-bank GRM. Dark outlines mark pairs above the positive $Q3 > 0.20$ demonstration screen.



Flagged pair	Q3	Item-total six	MILP six
DEMO_01 / DEMO_02	0.367	Both retained	Not both retained
DEMO_03 / DEMO_04	0.362	Both retained	Not both retained

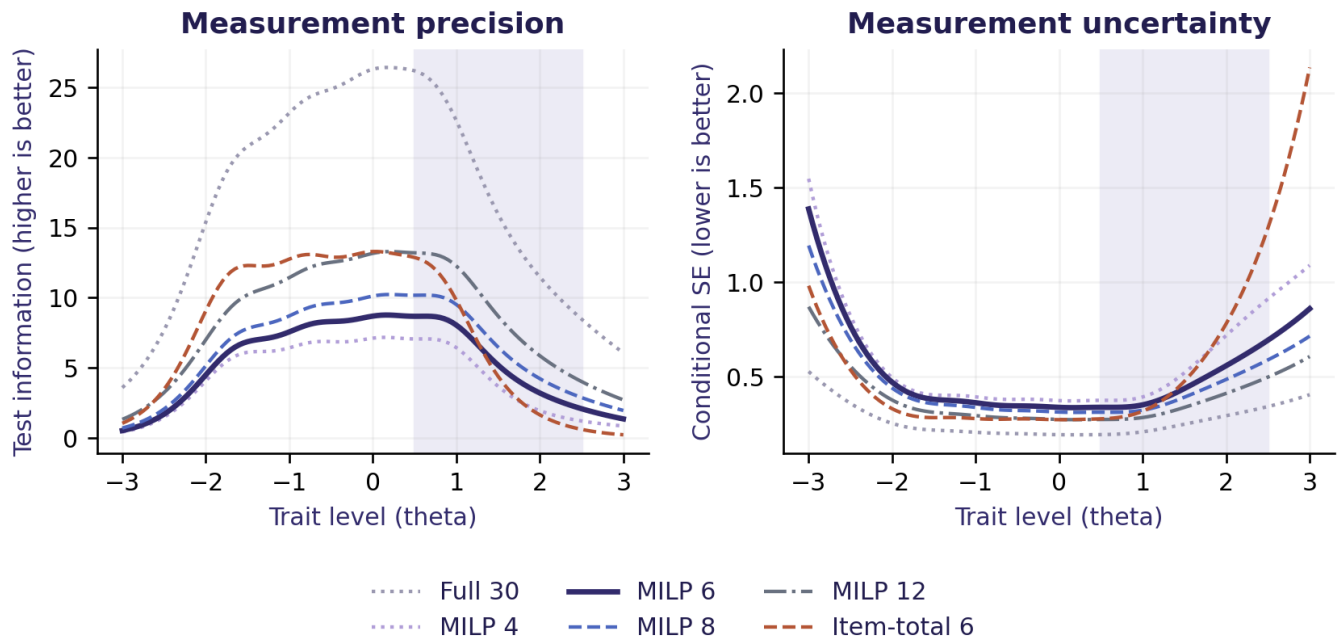
Content coverage	Anxiety	Anger	Depression
Full bank	10	10	10
Item-total six	5	1	0
MILP six	2	2	2

Two item pairs were deliberately made dependent by copying one generated response to the other with probability 0.45. The audit detects this simulation mechanism; it does not show that the actual IPIP wording is redundant in observed respondents. The comparator selects the six largest corrected item-total correlations, **not** the exhaustive alpha-maximizing subset. Positive $Q3 > 0.20$ is a workflow screen here, not a universal pass/fail rule.

03 / PRECISION ACROSS SCORE LEVELS

The shortest form is not always enough

Fitted test information (TIF) and conditional standard error (CSEM) for every candidate. Shading marks the illustrative target range, theta = 0.5 to 2.5.



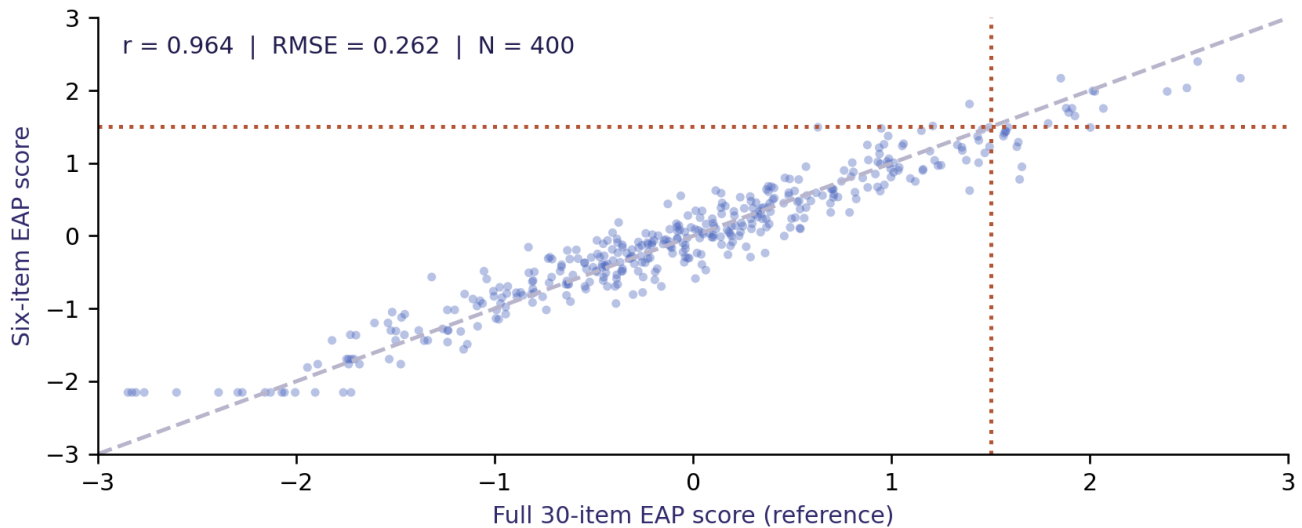
Form	Items	Mean target information	SE at theta 1.5
Full 30	30	16.90	0.250
MILP 4	4	4.06	0.522
MILP 6	6	5.47	0.439
MILP 8	8	6.70	0.393
MILP 12	12	8.88	0.339
Item-total 6	6	5.68	0.477

The MILP maximizes average target-zone information subject to length, minimum content counts and exclusion of flagged pairs. The six-item form retains two items per facet. No reading-grade constraint is asserted because no readability assessment was performed. The item-total comparator has slightly higher mean target information but retains both dependent pairs; its summed information assumes independence that the audit contradicts. Conditional $SE = 1 / \sqrt{\text{information}}$ is model-based, not an empirical error guarantee or marginal reliability estimate.

04 / HELD-OUT SCORE COMPARISON

High correlation can still miss cases

Holdout N = 400. Short-form EAP scores use frozen training-bank parameters; no short-form refitting or holdout-based item selection.



Six-item candidate: $r = 0.964$; $RMSE = 0.262$; agreement = 97.0%. At $\theta \geq 1.5$, it detects 13 of 23 full-bank reference-positive scores and misses 10.

Form	TP / FN	TN / FP	Sensitivity	Specificity
MILP 4	12 / 11	373 / 4	52.2%	98.9%
MILP 6	13 / 10	375 / 2	56.5%	99.5%
MILP 8	16 / 7	375 / 2	69.6%	99.5%
MILP 12	16 / 7	373 / 4	69.6%	98.9%
Item-total 6	15 / 8	372 / 5	65.2%	98.7%

Six-item Wilson 95% intervals: sensitivity 36.8% to 74.4%; specificity 98.1% to 99.9%. Only 23 reference-positive scores occur; uncertainty is substantial. The 30-item score contains the short-form items, so this correlation benefits from shared items. This is independent-person validation, not an independent-instrument criterion.

Interpretation: The cutoff is arbitrary and illustrative, not a validated clinical screening threshold. Sensitivity/specificity describe agreement with the full-bank score classification, not diagnosis. The six-item candidate is **not recommended as a clinical screener** on these results. Larger forms or revised optimization objectives may better suit high-sensitivity use.

05 / TRANSPARENT SCORING

A complete raw-score crosswalk

Six-item candidate: DEMO_02, DEMO_03, DEMO_14, DEMO_17, DEMO_22, DEMO_30. Each already-keyed item is scored 0-4; total range is 0-24.

The table computes $E[\theta | \text{raw sum}]$ from the frozen GRM by category-probability convolution, with a standard normal prior and quadrature from -6 to +6 (step 0.025). This is a model-based summed-score approximation; people with the same total can have different response-pattern EAP scores.

Raw sum	Theta EAP	Demo T-score	Posterior SD (T units)
0	-2.154	28.5	4.9
1	-1.677	33.2	4.0
2	-1.387	36.1	3.8
3	-1.153	38.5	3.7
4	-0.952	40.5	3.7
5	-0.769	42.3	3.7
6	-0.599	44.0	3.7
7	-0.439	45.6	3.7
8	-0.286	47.1	3.7
9	-0.139	48.6	3.7
10	0.004	50.0	3.7
11	0.145	51.4	3.7
12	0.284	52.8	3.7
13	0.422	54.2	3.7
14	0.562	55.6	3.8
15	0.704	57.0	3.8
16	0.850	58.5	3.9
17	0.999	60.0	4.0
18	1.154	61.5	4.1
19	1.314	63.1	4.2
20	1.490	64.9	4.4
21	1.668	66.7	4.6
22	1.875	68.8	4.9
23	2.065	70.7	5.1
24	2.398	74.0	5.7

Not normative: Demo T = 50 + 10 x theta. This linear transformation does not create population norms, PROMIS scores, clinical interpretations or an equivalent validated IPIP score. Posterior SD describes individual latent-score uncertainty conditional on fixed parameters; it is **not** the standard error of a group mean. Item-parameter uncertainty is not included.

A methods paragraph you can audit

Example wording for this simulated demonstration only. Replace it with the actual design and results before using it in a manuscript.

Responses were simulated for 1,200 individuals on a custom 30-item bank whose wording was drawn from the public-domain IPIP Anxiety, Anger and Depression facets. The unidimensional graded response model was fitted in mirt to the first 800 response records. Two pairs were deliberately made locally dependent by response copying (probability 0.45). Training-sample positive Yen Q3 values above 0.20 identified pairs subject to mutual-exclusion constraints. Binary linear programming maximized mean model-based information over theta 0.5-2.5 for lengths 4, 6, 8 and 12, with at least floor(length/3) items from each facet. A comparator retained the six highest corrected item-total correlations. Item parameters were frozen for EAP scoring of 400 held-out individuals, with unadministered items treated as missing. The six-item candidate correlated 0.964 with full-bank scores (RMSE 0.262); agreement at the illustrative theta 1.5 cutoff was 97.0%, with sensitivity 56.5% and specificity 99.5%. A summed-score EAP crosswalk used probability convolution and a standard normal prior. These simulated results do not establish clinical validity.

What has and has not been established

The GRM optimizer converged. SciPy independently reproduced the R item-information curves within 1e-8 and MILP objective values within 1e-7. This demonstration does not assess empirical dimensionality, subgroup DIF, readability, external validity, longitudinal stability or clinical utility. The full-bank fit intentionally includes dependence, so additivity of information and score uncertainty may be optimistic. A real engagement requires model-fit and construct review before selecting a final form.

Reproduce and inspect

Run generate_audit.R, then check_and_plot.py, then build_report.py from the repository root. The accompanying ZIP includes scripts, item source mapping, synthetic responses, fitted parameters, all candidate selections, curves, Q3 matrix, holdout scores, metrics, crosswalk, session information and checksums. R calibrates; Python cross-checks the frozen calibration and assembly objectives. The two scripts are not claimed to be independent end-to-end calibrators.

Sources and technical references

IPIP item key (N1 Anxiety, N2 Anger, N3 Depression)

<https://ipip.ori.org/newNEOKey.htm>

IPIP public-domain permission statement

<https://ipip.ori.org/newPermission.htm>

Chalmers (2012). mirt. Journal of Statistical Software, 48(6).

<https://doi.org/10.18637/jss.v048.i06>

mirt scoring and Q3 documentation

<https://philchalmers.github.io/mirt/reference/fscores.html>