

Psychometric and mathematical audit

TypologyOnline DISC-inspired personal reflection · Release 0.3.0 · 22 October 2026

1. Decision and intended interpretation

The defensible result is a transparent summary of a person's current self-reported answers to four broad behavioural item sets. The evidence supports some consistency and theoretically sensible construct correspondence, especially in the longer forms. It does not establish a precisely accurate, stable DISC type, four pure personality traits, or the probability that a personal interpretation is true.

The recommended version for personal use is the 48-item form. It gives each broad summary twelve items rather than four. The 16-item form remains an introduction, but its S and C numerical style scores are withheld in the user interface because their consistency is weak across three cohorts. The person still receives their actual S/C answers and reflection prompts. Medium and long numerical results are labelled descriptive answer summaries.

This report justifies the interpretations that the evidence can support and identifies those it cannot. A mathematical derivation can prove that a response was scored correctly. It cannot prove that a self-description is psychologically true, that a behaviour is stable, or that a questionnaire covers all relevant aspects of a person. Personal use makes careful, modest interpretation appropriate; it does not remove measurement error.

The practical offering can therefore be made in good faith as an **experimental, voluntary personal reflection**, with the limits visible before and after taking it. Marketing it as a validated DISC diagnosis, a scientifically established type, or a test with an accuracy percentage would not follow from this audit.

No independent psychometric reviewer, prospective website sample, observer study or test-retest sample has evaluated this instrument. This is a reproducible technical development audit, not a certification or peer-reviewed validation study.

2. What was independently checked

We froze the 48 original items, their form memberships, reverse keys and equal weights before examining the later response cohorts. We did not select alternative item subsets after seeing the replication coefficients. A local protocol documents this decision. It is not an independently registered preregistration; the first cohort's earlier standardized-alpha estimates were already known.

The audit checked the actual scoring equation, item identity across releases, missingness and pairwise sample sizes, raw-score versus standardized consistency, respondent-row bootstrap uncertainty, temporal replication, ordinal factor models, approximate model-based reliability, alternative split halves, non-overlapping construct anchors, and recorded age/gender sensitivity. The resulting changes concentrate on the interpretation and usefulness of the personal report.

The arithmetic was also checked against the independent psych R package implementation of raw alpha. All twelve cohort-by-long-scale comparisons agreed within $1e-10$. That cross-check verifies this calculation; it does not validate the intended personality interpretation.

The audit's strongest additional evidence is temporal replication in separate collection periods, rather than repeatedly splitting the same initial dataset. The same platform and measurement method remain common to all periods. We cannot establish that every record belongs to a different human across periods.

3. Data, provenance and analysis populations

The sources are original Harvard Dataverse SAPA releases, with public metadata identifying CC0 1.0 for the pinned versions. Exact input file IDs and SHA-256 hashes are archived. All 48 questionnaire items and all 36 separate anchor items were present in each of the three 696-item releases. Item matching preserved words; later dictionaries contain quotation/apostrophe serialization differences, which were normalized for matching. The displayed questionnaire keeps the verified original IPIP wording.

Cohort / period	DOI	Released rows	Selected-pair N range
A / Dec 2013-Jul 2014	10.7910/DVN/SD7SVE	23,679	314-964
B / Jul 2014-Dec 2015	10.7910/DVN/GU70EV	54,855	704-2,106
C / Dec 2015-Feb 2017	10.7910/DVN/TZJGAT	48,350	595-1,885

The analysed total is **126,884 released response records**, not that many people completing the new 48-item questionnaire. Unique RID values match the row count within each release. Source period boundaries and identifiers do not establish unique people across releases. The earlier C40K1U release was checked for coverage but excluded from this analysis because it lacks thirteen selected items.

SAPA used random subsets of a larger item bank and the same six response anchors. A person was not expected to answer all 696 questions. We retained missing responses, checked observed selected responses were integers from 1 through 6, and used available joint observations for each item pair. No missing answer was replaced by a neutral value. The analysis includes the released records as provided, without filtering on attractive score profiles or deleting identical-answer patterns.

The data paper describes the planned sampling as missing completely at random. Our available-case moment and pairwise-likelihood analyses rely on the relevant sampling/missingness assumptions. Random item allocation supports those assumptions; voluntary stopping and other participation effects are not automatically proved ignorable. No new MCAR test or causal argument about all participation mechanisms was established here. [1-4]

Age is stored in bins, including 18andUnder. An exact 18+ subset cannot be recovered. The older sensitivity subset therefore uses the identifiable 19+ bins. Gender sensitivity uses the two recorded source categories, male and female; this does not establish validity for unrepresented identities. These convenience samples are not population norms for the site's intended users.

Cohort C is a temporal replication for the item sets that were fixed before its responses were analysed. It is not a prospective validation of the revised website report, and its historical label 'Confirmatory Sample' is not itself confirmation of our DISC interpretations.

4. Item content and the meaning of the four summaries

The item assignments are DISC-inspired development hypotheses. Public-domain wording and a history of research with source inventories do not automatically validate a new regrouping of those items. The official IPIP source allows reuse; its reliability figures for other full scales are not inherited by this instrument. [5]

Summary	What its items actually cover	Main interpretation limits
D	Assertion, initiative, leading, perceived leadership confidence	Several items concern confidence/ability beliefs, control or task execution; those do not demonstrate skill.

Summary	What its items actually cover	Main interpretation limits
I	Social initiation, ease, warmth, group engagement, humour and influence confidence	Social ease, humour and persuasive self-belief are distinguishable; some items overlap D or S.
S	Patience/anger regulation, practical care, cooperation and keeping promises	A broad composite of several behaviours; routine preference is poorly represented. A single score can hide an uneven pattern.
C	Detail attention, planning, deliberation, order and standards	A broad composite rather than a pure analytical trait; not an intelligence or decision-quality measure.

The item-by-item appendix identifies specific ambiguity, context, redundancy and construct contamination. It is a development content audit, not independent expert adjudication or cognitive interviewing. Its purpose is to keep the interpretation honest and to identify the next research decisions.

We did not replace items merely to increase alpha. Repeatedly asking almost the same question can raise consistency while narrowing coverage. The original four-item S set, for example, combines patience, anger, helping and promises: this breadth helps explain why a short homogeneous score is difficult. The longer S set contains both patience and support clusters, and an almost opposite-wording promises pair. A high total alpha would not erase those distinctions.

No circular DISC structure or twelve profile categories was demonstrated. D and I overlap positively, and the four scores are not forced to sum to a constant. Treating the letters as four exclusive kinds of people, or converting them to another personality system, would add an unsupported interpretation.

5. Exact scoring: a mathematical specification

For person i and item j , let x_{ij} be the selected integer in $\{1,2,3,4,5,6\}$. Let $s_j = +1$ for a forward key and -1 for a reverse key. Define the keyed response y_{ij} as:

$$\begin{aligned}
 y_{ij} &= 3.5 + s_j * (x_{ij} - 3.5) \\
 &= x_{ij} && \text{for a forward-keyed item} \\
 &= 7 - x_{ij} && \text{for a reverse-keyed item.}
 \end{aligned}$$

The reversal is 7 minus the answer because there are six categories. Using 6 minus the answer, appropriate to a five-point example, would be an error here. Both original and keyed values remain in $[1,6]$.

For dimension d with k_d items, define its keyed mean M_{id} and 0-100 arithmetic summary Z_{id} :

$$\begin{aligned}
 M_{id} &= (1 / k_d) * \sum_j (y_{ij}) \\
 Z_{id} &= 100 * (M_{id} - 1) / (6 - 1) \\
 &= 20 * M_{id} - 20.
 \end{aligned}$$

Every question has equal weight. We calculate all four summaries separately. No demographic correction or adaptive weighting is applied. Release 0.3.0 adds a separate, model-estimated historical-reference percentile after scoring (Appendix C); it does not change the score or claim representative population norms. All items in the selected form must be answered before a result is computed. Unknown forms, missing/invalid answers and incompatible saved state are rejected or handled explicitly.

For four keyed values 6, 4, 5, 3, the sum is 18, $M = 4.5$ and $Z = 70$. An original answer of 2 to a reverse-keyed item contributes 5. All keyed answers of 1 produce 0 and all keyed answers of 6 produce 100. These endpoint cases are possible because opposite-keyed original responses differ.

The scoring has several properties that can be proved without assuming a personality model:

- Boundedness: because each keyed value is between 1 and 6, its average is between 1 and 6, hence Z lies between 0 and 100.
- Monotonicity: increasing an original forward response by one increases its dimension summary by $20/k_d$. Increasing an original reverse response by one decreases it by $20/k_d$. Other dimensions are unchanged.
- Granularity: one response step changes the summary by 5 points in a four-item scale, 2.5 in an eight-item scale and 1.6667 in a twelve-item scale. Displaying whole numbers does not create one-point psychological precision.
- Scale invariance: converting a sum to a mean or applying a positive affine transformation changes units, not reliability. If $X^* = aX + b$, $\text{Var}(X^*) = a^2 \text{Var}(X)$; the ratio of true to observed variance is unchanged.
- Non-composition: no equation constrains $D + I + S + C$ to equal 100. A pie chart would misrepresent this scoring.

Each form has three forward items for every reverse item. If a person chooses the same original answer c on every question, the keyed mean is $(3c + (7-c))/4$ and the displayed arithmetic summary is:

$$Z_{\text{constant}} = 10 * c + 15.$$

Thus identical original answers of 1 through 6 yield summaries of 25, 35, 45, 55, 65 and 75. Mixed keys do not by themselves cancel agreement response style. The application flags identical answering gently and permits review; it does not label the person dishonest or apply an unvalidated exclusion rule.

As another mathematical check, independent uniformly random answers have $E[x] = 3.5$ and $\text{Var}(x) = 35/12$. Reversal leaves that distribution unchanged, so $E[Z] = 50$ and $\text{Var}(Z) = 400 * (35/12) / k_d$. Random answers can therefore produce plausible-looking numerical summaries. A result's appearance is not evidence of its accuracy.

The JavaScript scorer uses exact small integer sums (maximum 72 per twelve-item scale); only division and display rounding introduce ordinary floating-point approximation. Exports preserve the unrounded calculations and original responses. The short S/C arithmetic values remain available in exports for transparent research reproduction, but their presentation metadata states they are not displayed as personal style scores.

6. Reliability aligned with the reported score

The earlier release reported standardized alpha from a Pearson correlation matrix. That was labelled correctly, but the application averages original keyed response values without standardizing each item. This audit therefore makes raw-score alpha primary and retains standardized alpha as a sensitivity statistic.

Let S be the estimated covariance matrix of the k keyed items and 1 a vector of k ones. The variance of their sum is $1' S 1$. Raw alpha is:

$$\text{alpha_raw} = k / (k-1) * [1 - \text{trace}(S) / (1' S 1)].$$

For the standardized version, let r_bar be the mean off-diagonal Pearson correlation:

$$\text{alpha_std} = k * r_bar / [1 + (k-1) * r_bar].$$

These statistics can differ when item variances differ. Equal weights on standardized variables are not identical to equal weights on the original categories. Neither coefficient is an 'accuracy percentage'. In classical test theory $X = T + E$ with $\text{Cov}(T, E) = 0$, reliability is $\text{Var}(T) / \text{Var}(X)$. Alpha estimates that ratio only under additional conditions; essential tau-equivalence and appropriate error assumptions matter. Correlated wording residuals and multidimensional content can make a simple interpretation misleading. [6-7]

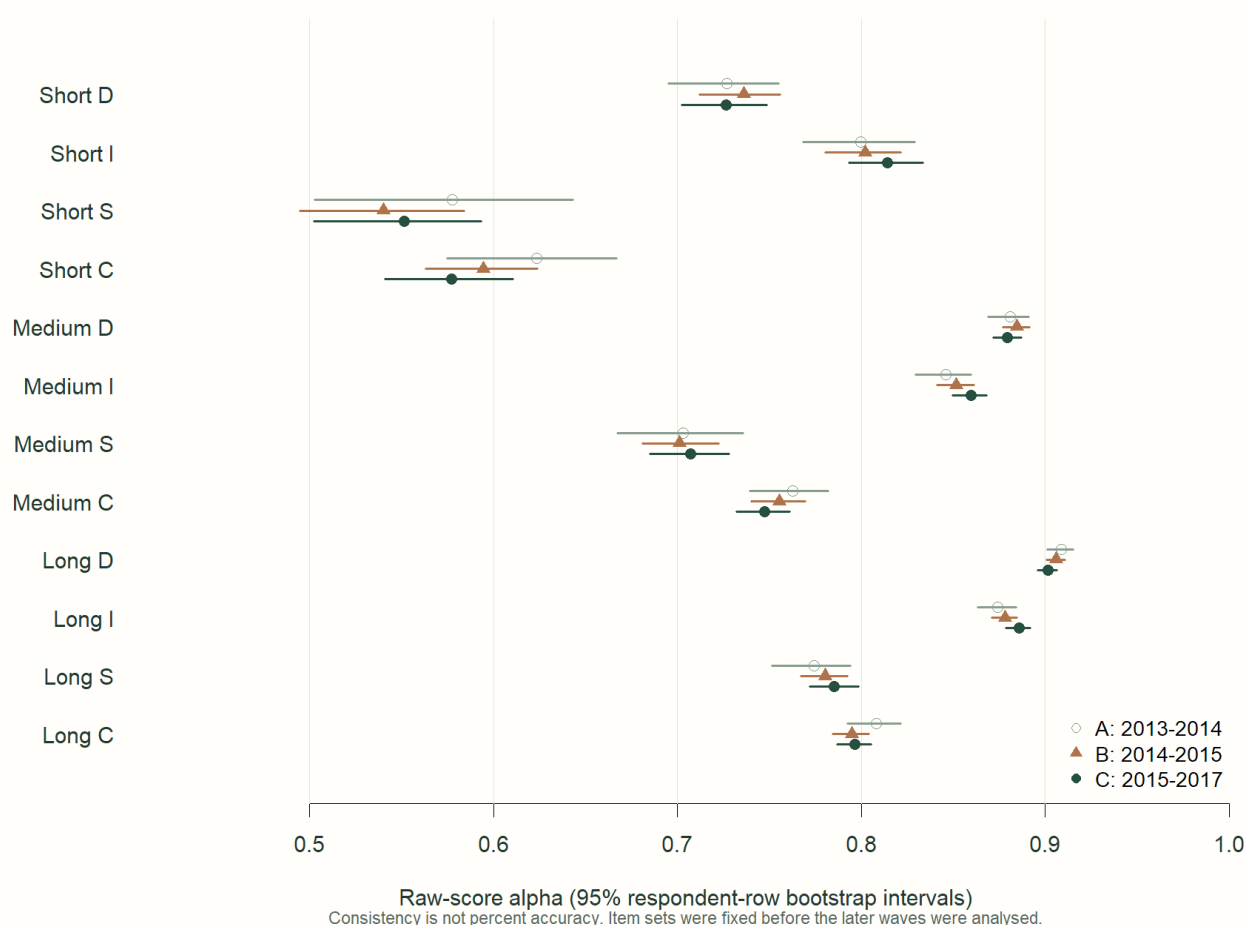
For each covariance entry S_{jl} we used the respondents with both items observed, their pair-specific means and a denominator of $n_{jl} - 1$. Item variances use the observations for that item. Pearson correlations also use pairwise available data. These moment estimators can be reasonable under matrix-sampling assumptions, but the finite-sample entries need not form a positive-semidefinite matrix. A single total N is not the N for every correlation.

The original cohort's full Pearson matrix had minimum eigenvalue -0.025634. The corresponding values in later cohorts are positive, about 0.089549 and 0.096862. Within-scale moment matrices used for the main coefficient calculations were checked separately. No smoothing or midpoint imputation was used to conceal the original non-positive matrix. The ordinal models use pairwise likelihood on the observed categories rather than pretending the original matrix is a complete-data covariance estimate.

6.1 Replicated raw-score consistency

Form / summary	A raw alpha	B raw alpha	C raw alpha	C 95% row-bootstrap interval
Short D	0.727	0.736	0.727	[0.702, 0.749]
Short I	0.800	0.802	0.814	[0.793, 0.834]
Short S	0.578	0.540	0.551	[0.502, 0.593]
Short C	0.624	0.594	0.577	[0.541, 0.610]
Medium D	0.881	0.885	0.879	[0.872, 0.887]
Medium I	0.846	0.852	0.860	[0.850, 0.868]
Medium S	0.703	0.701	0.707	[0.685, 0.728]
Medium C	0.763	0.756	0.747	[0.732, 0.761]
Long D	0.909	0.906	0.901	[0.896, 0.906]
Long I	0.874	0.878	0.886	[0.879, 0.892]
Long S	0.774	0.781	0.785	[0.772, 0.799]
Long C	0.808	0.795	0.796	[0.787, 0.805]

Consistency across three SAPA collection periods



Consistency estimates with respondent-row bootstrap intervals

The longer forms consistently show higher coefficients for these particular item sets. D and I long-form estimates are comparatively strong. The broader S/C long-form composites have more moderate estimates. The four-item S and C summaries remain weak across periods; a larger source dataset did not repair their low consistency.

The interval in the table describes uncertainty about a population-level consistency coefficient under the source sampling assumptions. It is not an interval for an individual's personality score, and it does not measure the probability that the report is correct.

6.2 How the intervals were calculated

For each cohort we drew 1,000 bootstrap samples, each containing the same number of complete respondent rows sampled with replacement. Each row retained all its observed and missing item values. We recalculated pairwise covariances/correlations, raw alpha, standardized alpha and the classical error illustration in each replicate. Percentile endpoints are the 0.025 and 0.975 quantiles, using R's default type-7 quantile interpolation. The random seed is 20261022.

Resampling whole rows preserves the empirical response associations and missingness patterns better than independently resampling cells or correlations. It still assumes rows are a suitable approximately independent resampling unit. Unknown repeat visitors or unmodelled selection can limit that assumption. The intervals include sampling variability of these moment estimates, not every source of bias. With 1,000 replicates, endpoints also have Monte Carlo variability. No normal-theory interval using the entire cohort N as every pair's N was substituted.

Full bootstrap distributions and all unrounded endpoints are supplied. The analysis reported no nonfinite raw-alpha bootstrap estimates in these twelve form/scale combinations in each cohort. The report does not equate the absence of numerical failures with psychological validity.

6.3 Alternative split halves and redundancy

We additionally enumerated all 462 unique equal six/six splits of each twelve-item set, calculated each pair of unit-weight half-score correlations from the covariance entries, and applied the equal-half Spearman-Brown expression $2r/(1+r)$. This is a descriptive sensitivity check. It shares the same matrix-sampling assumptions and is not an independent alternate-form validation.

Long summary	C minimum split coefficient	C median split coefficient	C maximum split coefficient
D	0.828	0.904	0.946
I	0.806	0.892	0.940
S	0.562	0.793	0.876
C	0.632	0.805	0.877

The wide S/C ranges show that the way content is divided between halves matters. A single alpha can hide that heterogeneity. More repetitive or more favourably divided items can yield a higher coefficient without improving interpretation of the full behavioural construct.

7. Structural validity: ordinal modelling and model-based reliability

Because the response categories are ordered, we fitted ordinal pairwise maximum-likelihood models with lavaan 0.7-3, ordered indicators, missing='pairwise', standardized latent variances, and no ordinary-ML test statistics. A one-common-factor model and the proposed four-correlated-factor model were fitted in all three cohorts; each twelve-item set also received an ordinal one-factor model in cohort C.

For an ordinal item, suppose an unobserved continuous response U_j is divided by thresholds $\tau_{j0}=-\infty$ through $\tau_{j6}=+\infty$. Category r occurs when $\tau_{j,r-1} < U_j \leq \tau_{jr}$. The model sets correlations between latent responses using its factor loadings and factor correlations. The probability of a pair of observed categories is the appropriate rectangle probability from a bivariate normal distribution.

$$CL(\theta) = \sum_i \sum_{(j < l, \text{ both observed for } i)} \log P_{\theta}(X_{ij} = r, X_{il} = s).$$

This is a composite pairwise likelihood, not a full joint likelihood for all 48 responses. Missing='pairwise' retains each available item pair rather than dropping every incomplete respondent. Cases with no selected observations provide no information and are omitted by the model; pairwise coverage warnings are preserved. Many respondents have no jointly observed pair for a particular comparison.

PML fit requires its own inferential treatment. We did not manufacture ordinary likelihood-ratio chi-square, AIC/BIC, CFI or RMSEA from a single nominal effective sample size. Standard errors and formal test statistics were disabled for these diagnostic models. The reported residual summary is explicitly descriptive:

$$RMSR_offdiag = \sqrt{\sum_{(j < l)} (r_{poly,jl} - r_{model,jl})^2 / m},$$

$$m = k*(k-1)/2.$$

Here r_{poly} is the empirical latent-response correlation estimate and r_{model} the fitted model correlation. This is called a descriptive off-diagonal SRMR/RMSR in the output. It is not a formal pass/fail certificate with an automatically applied universal threshold. [8]

Cohort	Model rows	One-factor residual RMS	Four-factor residual RMS	Four-factor max absolute residual
A	23,446	0.136	0.102	0.540
B	54,303	0.131	0.097	0.519
C	47,890	0.131	0.098	0.545

The intended four-factor models converged and returned admissible point estimates, but substantial systematic residual correlations remained. They reproduce the data better than a single common factor; that relative improvement is insufficient to claim a clean four-trait model. Software convergence means a numerical solution was found, not that the model is true.

In cohort C, fitting each item set as one ordinal factor gave these residual summaries:

Twelve-item summary	Ordinal one-factor residual RMS	Maximum absolute residual
D	0.058	0.164
I	0.089	0.344
S	0.126	0.553
C	0.118	0.509

The largest residual pairs in the full four-factor cohort-C model make the limitation concrete:

Item pair	Observed latent-response r	Model r	Residual
C02 / C06	0.648	0.103	0.545
S02 / S09	0.556	0.068	0.488
S04 / S10	0.800	0.330	0.470
D12 / C11	0.556	0.154	0.403

C02/C06 both concern rash action; S02/S09 both concern anger; S04/S10 are opposite-wording promises items; D12/C11 both concern carrying plans through. These within-theme and cross-summary relationships are too large to dismiss as a clean set of independent item errors. The detailed residual file includes every item pair.

7.1 Why omega does not fix an inadequate model

As a sensitivity analysis, we fitted a one-factor MINRES approximation to each long-form Pearson covariance structure. With a congeneric factor F of variance 1 and uncorrelated residuals, the unit-weight reliability expression is:

$$Y_j = \text{intercept}_j + \lambda_j F + \epsilon_j$$

$$\omega = \frac{(\sum_j \lambda_j)^2}{(\sum_j \lambda_j)^2 + \sum_j \theta_j}$$

For this calculation, standardized loadings from $\text{cov2cor}(S)$ were multiplied by the original item standard deviations; residual variances were $(1 - \lambda_j^2) \cdot \text{SD}_j^2$. This aligns the approximation with unit weights on original

response units. It remains a Pearson/continuous approximation, not a categorical observed-score reliability solution, and it depends on a defensible one-factor and local-independence model.

Long summary	C raw alpha	C approximate one-factor omega	Pearson one-factor residual RMS
D	0.901	0.905	0.060
I	0.886	0.893	0.083
S	0.785	0.783	0.121
C	0.796	0.797	0.106

These omega values are reported for completeness and reproducibility, not selected as larger or more impressive replacements for alpha. The residuals, especially for S/C, challenge a pure common-factor interpretation. Neither alpha nor this conditional omega is a calibrated statement that a given person's description is correct. Their agreement does not overcome model misspecification. A general-factor saturation or ordinal reliability result cannot be inferred simply by relabelling these numbers. [6-7]

8. Correspondence with separate construct anchors

We specified 36 non-overlapping IPIP items, six per anchor, for assertion, social ease, compassion, patience, orderliness and deliberation. Their wordings/keys were taken from official IPIP source keys and matched in each dataset before examining the later-wave results. No selected anchor is one of the 48 questionnaire items.

These are custom six-item non-overlapping research anchors, not the unchanged complete established source scales. We do not inherit their original names' published reliability figures or call them fully validated instruments in their reduced form. They provide a limited check on whether a new summary relates to the intended content without exact item overlap artificially creating the correlation.

Because complete paired scores are generally unavailable under the matrix-sampling design, the correlations are reconstructed from available-case covariance entries. With weight vectors u and v selecting the two non-overlapping composites:

$$r_{\text{composite}} = (u' S v) / \sqrt{(u' S u)(v' S v)}.$$

This estimates a correlation between hypothetical complete unit-weight composites under the moment/missingness assumptions. It is not an observed complete-questionnaire correlation for all records. The results are not disattenuated, and no bootstrap intervals or significance tests for these correlations are claimed in this release.

Long summary	Assertiveness	Social ease	Compassion	Patience	Orderliness	Deliberation
D	0.772	0.527	0.167	0.065	0.187	-0.079
I	0.673	0.819	0.407	0.190	0.055	-0.217
S	0.087	0.416	0.599	0.591	0.158	0.203
C	0.068	0.060	0.139	0.114	0.654	0.590

Content correspondence with separate IPIP anchor items

	Assertion	Social ease	Compassion	Patience	Order	Deliberation
D	0.77	0.53	0.17	0.06	0.19	-0.08
I	0.67	0.82	0.41	0.19	0.05	-0.22
S	0.09	0.42	0.60	0.59	0.16	0.20
C	0.07	0.06	0.14	0.11	0.65	0.59

Cohort C. Matrix-reconstructed raw-score correlations; no shared item IDs; same self-report method.

Non-overlapping construct-anchor correlations in cohort C

The patterns are theoretically sensible and replicate in the other periods. D relates most strongly to the assertion anchor; I to social ease; S to both compassion and patience; C to both order and deliberation. They support those narrow content interpretations more directly than they support a special DISC taxonomy.

All anchors and questionnaire items were self-reported on the same survey platform. Common response style, desirability and method variance can contribute to the correlations. Different item IDs remove exact item overlap, not shared method or shared wording themes. We have no independent observer, experimental behaviour, future outcome or commercial DISC criterion. A modest positive correspondence should not be converted into a 'percentage match' or proof of individual accuracy.

8.1 Distinctness and raw-score comparability

The reconstructed long D/I score correlation in cohort C is shown below with the other summary correlations. There is meaningful overlap, so four unrelated or orthogonal traits would be an inaccurate description.

Summary	D	I	S	C
D	1.000	0.657	0.207	0.243
I	0.657	1.000	0.373	0.047
S	0.207	0.373	1.000	0.336
C	0.243	0.047	0.336	1.000

The source means also differ between question sets, without establishing that their underlying psychological traits have a common metric:

Cohort / long-form mean	D	I	S	C
A	62.6	61.3	69.9	67.1
B	62.3	60.6	69.6	66.6
C	61.8	60.0	69.4	66.5

A person's largest raw number can therefore reflect item endorsement difficulty and content as well as their tendencies. Putting all four answers on a 0-100 display does not equate four constructs. We removed the most-endorsed/dominant-style headline rather than use convenience-sample means to manufacture calibrated type rankings. We do not supply representative general-population percentiles, pie-chart percentages or norm-based high/low categories. Release 0.3.0 separately adds explicitly model-estimated reference-sample midranks, with their assumptions and limits in Appendix C; those do not justify dominant-type classification. The IPIP project's own norms guidance explains why a reference population must be justified rather than assumed. [9]

9. Individual precision: what the mathematics allows

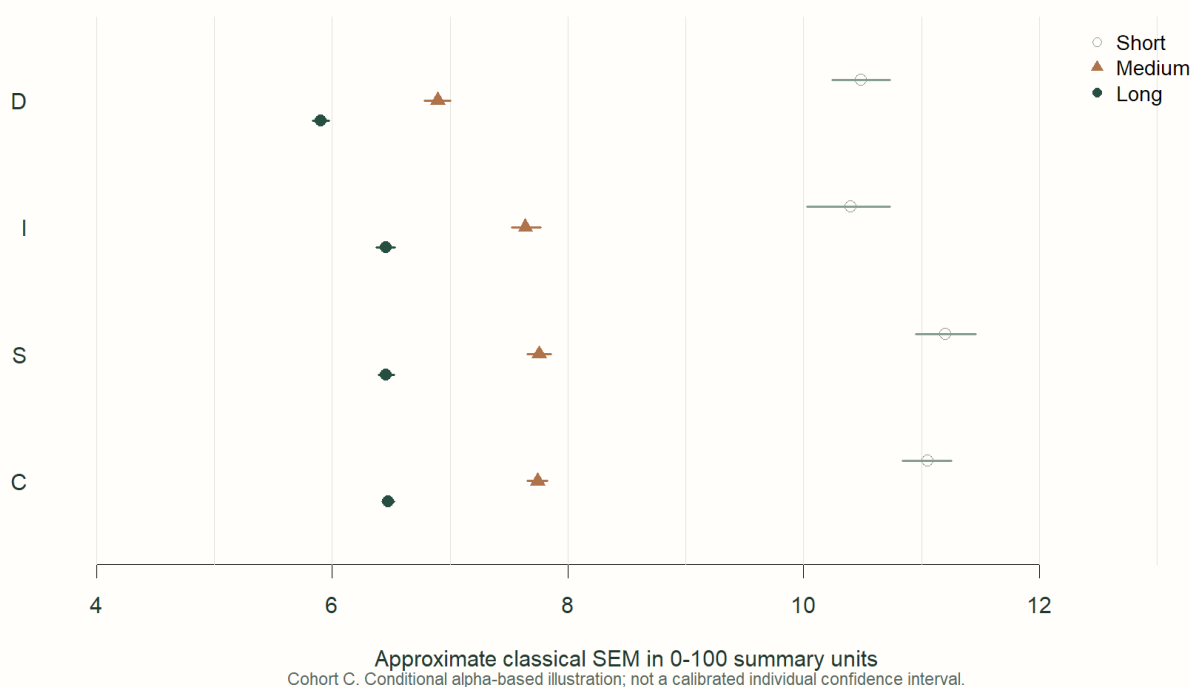
Under a classical model with an appropriate reliability coefficient ρ , the standard error of measurement is $SD_X \sqrt{1-\rho}$. Substituting raw alpha and the matrix-derived complete-score variance gives a useful **conditional error illustration**, in the displayed 0-100 units:

$$\begin{aligned} \text{Var}(M) &= (1' S 1) / k^2 \\ SD_Z &= (20/k) * \sqrt{1' S 1} \\ SEM_Z &= SD_Z * \sqrt{1-\alpha_{\text{raw}}}. \end{aligned}$$

The calculation estimates the error scale under classical assumptions about alpha, the covariance population and score interpretation. It does not calibrate an individual's conditional error, account for all bias or establish temporal stability. In a multidimensional composite, the distinction between dependable specific content and the intended broad trait is particularly important.

Form / cohort C	D approximate SEM	I approximate SEM	S approximate SEM	C approximate SEM
Short	10.49	10.40	11.21	11.05
Medium	6.89	7.64	7.76	7.74
Long	5.91	6.46	6.46	6.48

Longer forms reduce an illustrative error scale



Classical SEM illustration with bootstrap uncertainty

For a long-form summary with an illustrative SEM around 6.5 points, multiplying by 1.96 gives about 12.7 points. That arithmetic is sometimes used to motivate a classical approximate interval under additional normal/error assumptions. It is not enough to give every website visitor a truthful personal 95% confidence interval, particularly when the latent model is imperfect. We therefore do not show such an interval around a person's score.

If errors from two measurements were independent and of equal size, the error SD of their difference would be $\sqrt{2} \times \text{SEM}$. Comparing different summaries also requires their error covariance, which we have not established. A five-point gap cannot responsibly be treated as proof of a dominant trait. The application avoids that inference.

Larger source samples narrow uncertainty about an alpha estimate. They do not by themselves reduce the measurement error of a person answering only four or twelve questions. Source N and individual test length solve different problems. Increasing relevant item coverage can improve precision; changing content also changes the meaning of a summary. The nested forms are not equated alternate forms, and their shared items would inflate a naive agreement calculation.

The classical 'true score' is the expected score over a specified measurement process. It is not a philosophical true personality, a verified self-description or a promise of the same behaviour in all situations. No model here decomposed stable trait, temporary state, wording-specific variance and error using repeated observations. [6]

10. Sensitivity, missing evidence and limitations

Recorded age and binary-gender subset calculations are supplied for every form/summary. For the long cohort-C summaries, the following coefficients illustrate their scope:

Recorded subset	N	D	I	S	C
all	48,350	0.901	0.886	0.785	0.796

Recorded subset	N	D	I	S	C
age 19 plus	34,403	0.900	0.884	0.778	0.797
age 19 29	21,814	0.903	0.879	0.788	0.781
age 30 plus	12,589	0.892	0.888	0.756	0.815
male	17,705	0.900	0.891	0.777	0.812
female	30,645	0.902	0.883	0.795	0.787

This is a descriptive sensitivity analysis, not evidence of measurement invariance or fairness. Similar alphas do not prove equal factor loadings, thresholds, intercepts, conditional error or meaning across groups. Some source categories are unavailable or too coarsely grouped. No differential-item-functioning or multi-group invariance study was completed. We did not train demographic adjustments that could create false precision or stereotypes.

Important evidence is still missing:

- Response processes: no cognitive interviews show how current website users understand the phrases or the forced even-numbered response anchors.
- Prospective replication: no new participants have taken this exact presentation and received this report within a validation study.
- Stability: no test-retest observations establish how much a score changes after weeks or with context.
- External validity: no informant ratings or independent behaviour establish the accuracy of a person's self-view.
- Individual calibration: no validated conditional error model, meaningful personal change threshold, profile cut-offs or population reference distribution exists.
- Practical benefit: no experiment or follow-up shows that reading the report improves self-understanding or decisions.
- Independent review: this audit has not been adjudicated by an outside qualified panel or published peer review.

The factor-model limitations are findings, not merely future tasks. In particular, S/C should be read as transparent composites of multiple themes. Those limitations are made visible in the results and method panel rather than buried in a technical appendix.

11. Changes to the personal report and their justification

Reporting safeguard retained in 0.3.0	Reason	What it establishes
Long form becomes the default	More item coverage and repeatedly higher consistency than the brief sets	A reasoned default; no guarantee of individual accuracy.
Short S/C prominent numerical scores withheld	Raw alpha approximately .54-.59 for these four-item sets across cohorts	A conservative presentation policy; original answers and arithmetic exports remain available.
No most-endorsed/do minant-style headline	Raw summary means/difficulties differ, with no calibrated cross-scale trait metric	Avoids an unsupported type inference.

Reporting safeguard retained in 0.3.0	Reason	What it establishes
Examples quoted from actual answers	One most-aligned and one least-aligned item illustrate what went into each summary	Verifiable personal feedback instead of unsupported generated traits.
Every answer can be reviewed by dimension	Broad totals can hide contextual or facet differences	Lets users inspect the evidence for themselves.
D wording includes stated leadership confidence	Some source items explicitly ask about perceived ability	Accurately describes the content without assessing skill.
S/C wording makes their breadth explicit	Content review, residuals and anchor patterns show multiple themes	Narrows interpretation to what was actually asked.
Consistency evidence uses raw alpha with intervals	The application averages original units and estimates have sampling uncertainty	Matches the evidence to the arithmetic; not an accuracy percentage.

The example-selection algorithm sorts a dimension's items by keyed response, using item rank for ties, and shows one most-aligned and one least-aligned original statement. Both are labelled with the person's original response anchor. It does not invent an opposite wording for a reversed item or pretend the examples independently prove the summary. The full answer list prevents the two examples from becoming the only evidence the user can inspect.

Numbers are still calculated consistently in exports, including the brief S/C arithmetic summaries. Export presentation metadata explicitly states which numerical styles were withheld and why. This preserves research transparency while avoiding a stronger personal interpretation in the interface.

The privacy design remains local: no answer collection, account, analytics or external scoring service is embedded. Device saving is opt-in. Version 0.3.0 uses a separate versioned storage key; previous release data are not silently reinterpreted as a new result. The earlier release's ZIP is retained separately for traceability.

11.1 Defensible public wording

“Explore how you describe your everyday behaviour using four DISC-inspired question sets. The longer forms show replicated item consistency in historical SAPA survey data, but the results remain provisional self-report summaries. Review the individual answers, consider context, and use the prompts for personal reflection.”

Avoid claims such as “your true DISC type”, “90% accurate”, “scientifically proven personality”, “validated for everyone”, “this result predicts your behaviour”, or “a high S score means you are a better teammate”. They would go beyond the evidence actually tested.

12. Validity argument: claim, evidence and boundary

Claim the user might understand	Available support	Appropriate conclusion
My responses were scored by the stated rule	Explicit bounded formula, reverse-key checks, independent recalculation and exports	Strong support for arithmetic fidelity, conditional on correct input.
The report faithfully shows what I answered	Direct original-item quotes and response anchors; all answers reviewable	A verifiable reflection of supplied responses.

Claim the user might understand	Available support	Appropriate conclusion
Long summaries have some repeatable item consistency	Three cohorts, raw/standardized alpha, row-bootstrap intervals and split diagnostics	Supported for the historical source samples under stated assumptions.
These themes relate to assertion, sociability, support/patience and detail/deliberation	Non-overlapping IPIP anchor correspondence and content mapping	Limited support for broad content alignment; shared-method limits remain.
Each number measures one pure DISC trait	Substantial residual correlations, broad S/C content and overlapping D/I	Not established; use descriptive composite interpretations.
This precisely and truthfully defines my stable personality	No website prospective study, informants, repeated observations or calibrated individual error	Not established. The report must invite checking against experience.
Reading the report benefits my life	Plausible reflection prompts, inspectable answers, low-stakes context	Potential utility, not demonstrated outcome effectiveness.

The intended claim is narrower than a conventional diagnostic personality profile: “this is how you answered these behavioural descriptions, summarized transparently”. The consistency and anchor evidence make that summary more considered than an arbitrary quiz. The structural failures and missing individual evidence define its limits. This intended-use reasoning follows the Standards' emphasis on evidence for specified interpretations rather than a universal label of 'test validity'. [10-11]

13. Next evidence that would change the conclusion

The highest-value next step is a prospective study of the full bank and report, with cognitive interviewing first. Include independent content review, a varied volunteer sample, separate observer or behavioural information where practical, and a retest subset after a predeclared interval such as two to four weeks. Record the setting and relevant changes rather than treating every change as error.

For design planning, a Fisher-z interval for a correlation uses approximately $SE[z] = 1/\sqrt{n-3}$. At $n=100$ and an observed $r=.80$, the illustrative 95% interval is about .72-.86; at $n=200$ it is about .74-.85. These calculations show that even moderate retest samples leave uncertainty. They are planning illustrations, not evidence collected for this test. A suitable sample size depends on desired precision, model complexity, missingness, heterogeneity and the intended comparisons.

If revised items or shorter forms are developed later, select them using a development sample and lock their keys before a separate replication. Do not tune repeatedly on the current cohort C and continue calling it an untouched confirmation sample. Evaluate the report interpretation as well as the coefficients. A stronger broad construct may require clearer item wording, more balanced coverage or separate facet reporting; a higher alpha alone is not the development objective.

No finite dataset will prove universally truthful personality descriptions. The appropriate improvement process is to narrow claims to demonstrated evidence, identify counterexamples and revise transparently. For this release, the tangible improvements are replicated consistency estimates, explicit structural findings and feedback that stays close to the user's own answers.

14. Reproducibility and numerical files

The self-contained test is index.html. This report is available as psychometric-report.html, a PDF under reports/, and editable Markdown under docs/. The public audit directory contains source metadata, anchor wordings, the local protocol, all point estimates, bootstrap distributions, pairwise matrices/counts, ordinal model summaries and item/residual diagnostics. Raw respondent files, installed R packages, RDS analysis inputs and full fitted objects are kept outside the deliverable.

```
python scripts/download_audit_data.py --cache /your/research-cache
# Install R packages psych, lavaan and jsonlite in an analysis library.
Rscript scripts/audit_statistics.R /path/to/disc-studio /your/research-cache
Rscript scripts/audit_models.R /path/to/disc-studio /your/research-cache
Rscript scripts/extra_diagnostics.R /path/to/disc-studio /your/research-cache
Rscript scripts/plot_audit.R /path/to/disc-studio
python scripts/write_psychometric_report.py
python scripts/render_psychometric_report.py
python scripts/build.py
node scripts/check.js
```

An optional third R-script argument supplies a separate R library directory. The downloader checks pinned hashes. Numerical functions use psych 2.6.9 and lavaan 0.7-3 in this audit; R is 4.5.3. Estimation results can differ with software updates or optimizer tolerances, so record versions and inspect warnings rather than silently overwriting conclusions. The scripts preserve warnings and model convergence information.

The browser application's consistency calculations do not run R or fetch research data. All necessary questions, score rules and concise evidence are embedded at build time. A person can run the delivered HTML without a server or analysis installation. Browser-tool restrictions limited direct file-URL QA; functional testing uses a local static preview and source checks for external runtime calls.

The current functional verification record is separate from this psychometric report. Software checks cannot be counted as respondents, reliability replications or validity evidence. They establish that the implemented rule and presentation behave as specified. Statistical evidence and interface verification answer different questions.

15. References and primary sources

- [1] Condon, D. M., & Revelle, W. (2015). [Selected personality data from the SAPA-Project: 08Dec2013 to 26Jul2014](#). Harvard Dataverse, pinned release 5.0, response file 11016303.
- [2] Condon, D. M., Roney, E., & Revelle, W. (2017). [Selected personality data from the SAPA-Project: 26Jul2014 to 22Dec2015](#). Harvard Dataverse, pinned release 3.0, response file 11016300.
- [3] Condon, D. M., Roney, E., & Revelle, W. (2017). [Selected personality data from the SAPA-Project: 22Dec2015 to 07Feb2017](#). Harvard Dataverse, pinned release 3.0, response file 11016294.
- [4] Condon, D. M., Roney, E., & Revelle, W. (2017). [A SAPA Project Update: On the Structure of phrased Self-Report Personality Items](#). Journal of Open Psychology Data, 5(1), 3. Collection design and response format.
- [5] [Official IPIP repository](#), [alphabetical wording bank](#), [BFAS source keys](#), [preliminary NEO facet keys](#), and [preliminary HEXACO keys](#). Wording/key provenance; public-domain policy. Source-scale coefficients are not inherited.
- [6] Revelle, W., & Condon, D. M. (2019). [Reliability from alpha to omega: A tutorial](#). Psychological Assessment, 31, 1395-1411; doi:10.1037/pas0000754. Reliability distinctions and conditional classical error concepts.
- [7] Flora, D. B. (2020). [Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right?](#). Advances in Methods and Practices in Psychological Science; doi:10.1177/2515245920951747. Model choice matters for interpreting omega.

- [8] Rosseel and colleagues, [lavaan categorical-data documentation](#) and [pairwise likelihood tutorial](#). Software-method documentation; current version recorded above.
- [9] IPIP, [Information about Norms and Data Sets](#). Relevance of a justified reference population.
- [10] AERA, APA, & NCME (2014). [Standards for Educational and Psychological Testing](#). Standards 1.0-1.3 motivate specified-use interpretation and explicit limits; no endorsement is claimed.
- [11] EFPA (2025). [Test Review Model](#). Review framework only; this project has no EFPA approval.

Appendix A. Item-by-item content audit

The wording below is public-domain IPIP material. Keys are the project's provisional DISC-inspired keys, not a claim that a source inventory validated this assignment. 'Reverse' means 7 minus the original response. The notes identify plausible interpretation limits that must be considered alongside the statistical diagnostics. Full IPIP and SAPA identifiers, form memberships and numerical item diagnostics are supplied separately.

ID / key	Statement after I...	Content/interpretation concern
D01 / forward	Take charge.	Taking charge depends on role and opportunity; initiative is not leadership quality.
D02 / reverse	Hold back my opinions.	Holding an opinion back can reflect timing, culture or strategy rather than low assertion.
D03 / forward	Am the first to act.	Being first to act can reflect activity or impulsivity as well as initiative.
D04 / forward	Am able to stand up for myself.	Standing up for oneself captures boundaries/confidence and may overlap social anxiety.
D05 / forward	Try to lead others.	Trying to lead reflects inclination in a context, not effectiveness or entitlement to lead.
D06 / reverse	Wait for others to lead the way.	Waiting for expertise can be sensible; reversal is a tendency assumption, not a value judgement.
D07 / forward	Take control of things.	Taking control is ambiguous about situation and can be interpreted coercively.
D08 / forward	Have leadership abilities.	An explicit perceived-ability item: it measures a belief about skill, not demonstrated skill.
D09 / forward	See myself as a good leader.	Self-rated leadership quality overlaps D08 and self-confidence; it is not independent evidence of ability.
D10 / reverse	Keep in the background.	Staying in the background can reflect familiarity, social ease or context, overlapping I.
D11 / forward	Seek to influence others.	Seeking influence overlaps I/persuasion and does not identify the method or effect.
D12 / forward	Turn plans into actions.	Executing plans touches industriousness; cross-summary association with C11 is a model concern.
I01 / forward	Start conversations.	Conversation initiation is relevant but depends on context, language and opportunity.
I02 / reverse	Am quiet around strangers.	Quietness with strangers need not imply quietness or low engagement with familiar people.
I03 / forward	Involve others in what I am doing.	Involving others also concerns collaboration/support, overlapping S.
I04 / forward	Make friends easily.	Friend-making self-belief depends on opportunities and social setting.
I05 / forward	Know how to captivate people.	Captivating people is an ability belief and overlaps assertion/influence in D.
I06 / reverse	Rarely enjoy being with people.	Social enjoyment can vary with mood, fatigue and circumstances rather than a stable trait.

ID / key	Statement after I...	Content/interpretation concern
I07 / forward	Warm up quickly to others.	Warming to others includes friendliness and warmth, overlapping S content.
I08 / forward	Amuse my friends.	Humour with friends is culturally/contextually specific; it is not all social expression.
I09 / forward	Feel comfortable around people.	Comfort around people partly reflects anxiety or confidence, not just engagement.
I10 / reverse	Seldom joke around.	A serious person can be socially engaged; low joking is an imperfect reverse indicator.
I11 / forward	Enjoy being part of a group.	Enjoying a group captures affiliation, which is not identical to social influence.
I12 / forward	Have a natural talent for influencing people.	A perceived talent item, overlapping D; cannot establish objective persuasiveness.
S01 / forward	Am usually a patient person.	A broad self-label for patience, susceptible to desirable self-description.
S02 / reverse	Lose my temper.	Temper concerns anger/emotional regulation; its content clusters with S09.
S03 / forward	Love to help others.	Helping measures prosocial inclination, not routine preference or consistency.
S04 / forward	Keep my promises.	Promises concern dutifulness and strongly overlap the reversed S10 wording.
S05 / forward	Value cooperation over competition.	Cooperation versus competition is a value preference, not observed support skill.
S06 / reverse	Take no time for others.	Time for others depends on workload, caring responsibilities and opportunity.
S07 / forward	Anticipate the needs of others.	Anticipating needs involves empathy and situational knowledge, not patience alone.
S08 / forward	Accept people as they are.	Acceptance reflects tolerance/agreeableness rather than stability or dependability.
S09 / forward	Rarely feel angry with people.	Low anger is related to S02 but is not interchangeable with caring or reliability.
S10 / reverse	Break my promises.	Opposite wording of promises adds redundancy and a shared-content residual with S04.
S11 / forward	Think of others first.	Putting others first can include self-sacrifice; higher endorsement is not inherently healthier.
S12 / forward	Remain calm under pressure.	Calm under pressure concerns emotional composure; not the same as support/cooperation.
C01 / forward	Pay attention to details.	Detail attention is self-reported and partly role-dependent.
C02 / reverse	Jump into things without thinking.	Unthinking action is impulsivity; its strong residual pair with C06 indicates a distinct cluster.
C03 / forward	Do things according to a plan.	Planning is relevant but can be adaptive or constrained by the environment.

ID / key	Statement after I...	Content/interpretation concern
C04 / forward	Avoid mistakes.	Avoiding mistakes is very broadly desirable and may involve anxiety rather than method.
C05 / forward	Have an eye for detail.	Eye for detail is a perceived ability and overlaps C01; no objective detail task was given.
C06 / reverse	Make rash decisions.	Rashness is a self-evaluation of decisions; it strongly overlaps reversed C02 content.
C07 / forward	Like order.	Liking order is a preference, not evidence of better work or reasoning.
C08 / forward	Choose my words with care.	Careful words may reflect tact or caution in communication, not analytical accuracy.
C09 / forward	Want every detail taken care of.	Every detail being covered can reflect high standards, control or perfectionism.
C10 / reverse	Am not bothered by disorder.	Tolerance of disorder is not identical to low organizational ability.
C11 / forward	Make plans and stick to them.	Sticking to plans adds persistence; it overlaps execution content in D12.
C12 / forward	Demand quality.	Demanding quality concerns standards; it does not demonstrate quality or precision.

Appendix B. Accuracy claims separated from arithmetic certainty

Mathematical correctness: given valid responses and the declared keys, the score can be recomputed exactly up to display rounding. This is the strongest kind of certainty established here.

Psychometric estimation: coefficients, model parameters and covariance-based correspondences are estimates from particular historical convenience samples. They have sampling uncertainty, assumptions and possible bias.

Individual interpretation: we do not know the probability that a person's report matches their actual behaviour or self-understanding. That would require an agreed accuracy criterion and observations not present in these datasets. No mathematical manipulation of alpha supplies that missing criterion.

Practical usefulness: inspectable answers and concrete reflection prompts provide a plausible mechanism for a useful conversation. Whether they improve outcomes is an empirical question that remains open. The honest offering is an aid to reflection whose evidence and limitations can be inspected, rather than a promised discovery of a person's definitive identity.

Appendix C. Estimated reference percentiles added in release 0.3.0

C1. Decision, target and limits

At the user's request we add a separate approximate comparison to a named historical reference group. The reference is the released SAPA cohort C, December 2015-February 2017 (DOI 10.7910/DVN/TZJGAT; 48,350 records). It includes the source's recorded ages and genders as released, without demographic weighting or matching. The target is the distribution of this exact questionnaire composite under a specified statistical reconstruction of those records, not the general population, the current website audience, a pure latent trait, or people of the user's age/gender/country.

The source is self-selected, historical and matrix-sampled. There are almost no complete eight- or twelve-item scores per theme, so direct empirical complete-form norming is not possible. Averaging whichever items a source respondent happened to answer would compare different tests with different score precision. We explicitly do not do that. We also do not silently substitute a normal CDF based only on a mean and standard deviation.

Instead, these are **model-estimated reference percentiles**. Their assumptions are an additional limitation, not evidence that the new questionnaire is validated. We show them as approximate, separately from the unchanged score, and name the reference and modelling on the result page. The short S/C percentile is withheld with its numerical style score. A percentile cannot fix an unreliable or broad composite. No dominant style, norm-based high/low category or claim of improved personality accuracy is added.

C2. Why a distributional model is necessary

Cohort	Form	Complete D	Complete I	Complete S	Complete C
A	short	43	34	44	39
A	medium	0	0	1	0
A	long	0	0	0	0
B	short	101	76	72	72
B	medium	2	0	1	1
B	long	0	0	0	0
C	short	84	55	50	75
C	medium	4	1	0	1
C	long	1	0	0	0

Each count is complete for that theme/form only; it is not complete across the full questionnaire. Counts expose why the total released N must not be described as that many complete-test norm scores. Item-pair sample sizes and missingness qualifications remain those in Section 3.

For each cohort and each twelve-item theme, estimate six-category marginal probabilities from all its observed keyed answers. Reverse coding occurs before modelling. Let $p[j,c]$ be the observed probability of category c and let $t[j,c] = \Phi^{-1}(\sum_{h \leq c} p[j,h])$ for $h \leq c$, $c=1,\dots,5$, with $t[j,0]=-\infty$ and $t[j,6]=\infty$. Φ is the standard normal cumulative distribution.

Let R be the unrestricted empirical twelve-by-twelve polychoric correlation matrix for those keyed items, extracted from lavaan's sample statistics in the archived ordinal analysis. It is the empirical ordinal matrix, not the fitted

four-factor covariance matrix. We impose no single-trait, four-factor, local-independence or independent-item restrictions. Every within-theme pair has its own relationship. All twelve cohort/theme matrices were positive definite; the script asserts that condition and never silently repairs or smooths a matrix.

Assume U is multivariate normal with zero means, unit variances and correlation R . Generate category $Y[j]=c$ when $t[j,c-1] < U[j] \leq t[j,c]$. This ordinal Gaussian-copula construction reproduces each estimated six-category marginal in expectation while encoding pairwise dependence. Its multivariate Gaussian dependence is a substantive assumption, not an established property of human personality responses. Flexible pairwise correlations do not remove higher-order or tail-dependence assumptions. Non-normal latent dependence can alter polychoric estimates and the resulting percentiles.

For each cohort/theme we draw 500,000 twelve-item vectors using R's normal generator, Cholesky decomposition and seed 20261023. The script uses R 4.5.3 and lavaan 0.7-3 from the main audit. The short and medium sums are the first four and eight items of those same vectors; long uses all twelve. Thus nested forms share one coherent within-theme model. Each form retains its own score distribution and lookup. No synthetic vector is represented as a real participant or added to the cohort N .

For a valid keyed sum s of k items, the reported reference midrank is

$$P(s) = 100 \times [\text{Pr_model}(S < s) + 0.5 \times \text{Pr_model}(S = s)].$$

Simulation replaces each probability with the corresponding relative frequency. In a count table with counts $n[j]$ for $j=k, \dots, 6k$, the lookup at index $s-k$ is $100 \times [(\text{sum of } n[j] \text{ for } j < s) + n[s]/2] / 500000$. This makes ties explicit. Since the /100 score is a strictly increasing affine function of the sum, ranking either quantity before display rounding gives the same midrank. The application uses the exact integer sum; it never looks up a rounded /100 display. Different forms and themes are not interchangeable tables.

The result is not the strict proportion below: half the modelled ties is included. For example, if 60% fall below and 10% tie, the midrank is 65, not 60 or 70. The display says About 65th and describes tied-score handling in its explanation. Estimates under 1 or over 99 are grouped as Below 1st or Above 99th to avoid unsupported tail precision. Internal/export decimals are reproducibility data, not claims of psychological precision.

C3. Distribution checks and temporal sensitivity

The next table compares the cohort C simulation with the complete-score means and standard deviations reconstructed directly from available item means/covariances. Matching moments is a diagnostic, not validation of the full distribution. Means and SDs are on the /100 score scale. The final column is the maximum difference, across all attainable sums, between the largest and smallest model midrank among cohorts A, B and C.

Form/theme	Mean: moment / model	SD: moment / model	Max cohort gap (points)
short/D	61.80 / 61.79	20.05 / 19.95	2.40
short/I	52.97 / 52.92	24.13 / 23.58	2.37
short/S	70.17 / 70.13	16.73 / 16.53	0.68
short/C	66.46 / 66.46	17.00 / 17.17	1.91
medium/D	62.88 / 62.87	19.86 / 19.66	1.28
medium/I	58.30 / 58.25	20.38 / 20.03	3.04
medium/S	71.72 / 71.70	14.33 / 14.14	1.60
medium/C	66.99 / 67.00	15.40 / 15.63	1.44

Form/theme	Mean: moment / model	SD: moment / model	Max cohort gap (points)
long/D	61.78 / 61.77	18.81 / 18.60	1.88
long/I	60.00 / 59.95	19.11 / 18.76	3.01
long/S	69.38 / 69.36	13.94 / 13.74	1.58
long/C	66.51 / 66.52	14.35 / 14.52	2.13

Historical agreement is encouraging within this survey platform. It does not measure sampling bias relative to the general population, model misspecification, current website transport, or individual score instability. The export's historical-cohort range is the minimum and maximum of these three model midranks at the supplied sum; it is a sensitivity range, not a 95% interval or a personal confidence interval. No percentile-table respondent bootstrap or full uncertainty propagation was completed.

As a limited out-of-period calibration check, compare the C-model cumulative distribution for the first four exact items with actual complete four-item sums pooled from earlier cohorts A/B. Earlier rows used for this check do not fit the C model, but source platform and participation mechanisms remain shared. The table shows the largest absolute empirical-CDF versus model-CDF discrepancy at any score threshold. Its small N reflects actual complete four-item records, not total cohort records.

Theme	Earlier complete N	Maximum CDF gap	95% DKW halfwidth
D	144	6.53 points	11.32 points
I	110	6.42 points	12.95 points
S	116	8.04 points	12.61 points
C	111	3.15 points	12.89 points

The Dvoretzky-Kiefer-Wolfowitz bound is $\epsilon = \sqrt{\log(2/0.05)/(2n)}$; under independent, identically distributed complete-case records it gives a simultaneous 95% band for that complete-case empirical CDF. It is a per-theme bound, not familywise across four themes. Model uncertainty and source-group differences are not included. The model curves fall within these broad bands, but this is not a precise calibration guarantee or a test proving the model true. Complete-case volunteers can differ from those who stopped earlier; the same source participation caveat remains.

Most importantly, four-item checks cannot validate eight- or twelve-item score distributions. Their higher-order joint dependence and tails remain assumptions. The selected source data do not identify those full-score distributions without distributional restrictions. Future complete-form samples must evaluate calibration before stronger normative wording is considered.

Conditional on a fixed fitted model, Monte Carlo uncertainty is small: a midrank is the average of independent variables taking values 0, 0.5 or 1, multiplied by 100, so its standard error is at most $50/\sqrt{500000}$, about 0.071 percentile points. This bounds simulation noise only. It does not bound parameter error, copula error, convenience-sample bias, demographic transport or the individual's measurement error. Those larger unknowns motivate approximate whole-number reporting and the visible experimental label.

C4. Reproduction, implementation and defensible wording

Run `reference_percentiles.R` after the main statistical/model scripts, with the same `root/cache/library` arguments. It reads the already reverse-coded source input and empirical ordinal sample matrices. The deliverable archives aggregate category probabilities, thresholds, ordinal matrices, simulation count distributions, source-completeness counts, moment checks and limited calibration counts in `research/audit/reference-percentile-analysis.json`. Raw participant rows and fitted RDS files remain outside the deliverable. The canonical browser table is `data/reference-percentiles.json`, embedded in `index.html`; the browser makes no network request.

Automated checks compare every lookup with the archived probability counts, verify monotonicity and bounds, check form-specific indexing, embedded/source equality, withheld short S/C values, and metadata. The JSON export includes the unrounded estimate, method, reference DOI and period, plus historical sensitivity. Original response scoring remains exactly as in Section 5.

Defensible wording: Estimated reference percentile, compared with historical SAPA volunteers; approximate model estimate. Unsupported wording: percentile in the general population, scientifically accurate personality percentile, age-adjusted rank, or observed ranking among 48,350 complete tests. A representative sample alone would still require evidence for this instrument and its intended use.

Methodological context: [Grønneberg, Moss and Foldnes, Partial Identification of Latent Correlations with Ordinal Data](#) explains why latent-correlation conclusions depend on distributional assumptions. The paper does not validate this instrument or this reconstruction. [Official IPIP norms guidance](#) cautions that a reference sample must be justified for the intended comparison. Here the comparison is deliberately restricted to a named historical convenience sample. A dedicated complete-form prospective calibration and broader sampling study remain the priorities for improving it.