ASSESSMENT HEALTH CHECK

Independent Psychometric Review

Manager Adaptability Profile — MAP-24

Prepared for	Instrument	Sample
Northbridge Talent Labs	**MAP-24 (24 items, 4 scales)**	**480 respondents**

Demonstration report — illustrative synthetic data. Northbridge Talent Labs is a fictional demonstration organisation; the MAP-24 is a fictional assessment.

Overall assessment

AMBER

Promising foundation — targeted revisions recommended before higher-stakes use.

- ✓ Four-factor structure broadly supported (CFI .94, RMSEA .063); a single dimension fits far worse.
- ✓ Three of four scales show strong internal consistency (ω .82–.88).
- △ Feedback Responsiveness is usable but weaker (ω .77) and needs refinement.
- △ Three item-level issues merit targeted review: RF6, the APS5/APS6 pair, and FR4.
- ✓ Scores broadly comparable across the two demonstration groups at scale level.
- ✕ The proposed “below 60/100 = development needed” interpretation has no empirical basis.

480
Respondents, complete data on all 24 items

4
Distinct dimensions supported; one overall dimension rejected

.77–.88
Scale reliability range (omega)

3
Item-level issues requiring targeted review (four items)

The MAP-24 has a credible psychometric foundation for developmental use. The intended four-dimensional structure is substantially better supported than a single overall dimension, and most items perform well. However, one weak Reflective Flexibility item, redundancy within Adaptive Problem Solving, weaker Feedback Responsiveness reliability and a localised subgroup difference on one item should be addressed. The current 60/100 interpretation is not supported by validation evidence and should not be used as a decision threshold.

Where a reasonable non-specialist reading could have gone wrong

“Overall alpha is .91, so one total score is fine”

A high alpha only says the 24 items hang together. The single-factor model fits poorly (CFI .73). Reliability does not establish unidimensionality. Page 2.

“RF6 lowers alpha, so delete it”

Dropping RF6 would raise Reflective Flexibility alpha from .81 to .84 — but it is the only item covering one strand of the construct definition. Review the wording first. Page 3.

“One item differs between groups, so the test is biased”

The difference on FR4 is real but small (ΔR^2 = .012); scale-level comparability holds. It needs review, not alarm. Page 4.

Structure and reliability

Does the MAP-24 measure four things, one thing, or something in between? A confirmatory factor model treating responses as ordered categories was fitted for the intended four-scale structure and compared against a single overall dimension.

Adaptive Problem Solving	
.88 omega	
Alpha	**.87**
Mean loading	**.76**
Lowest loading	**.70**
Scale mean (1–7)	**5.10**

Feedback Responsiveness	
.77 omega	
Alpha	**.78**
Mean loading	**.62**
Lowest loading	**.54**
Scale mean (1–7)	**4.84**

Change Self-Efficacy	
.88 omega	
Alpha	**.88**
Mean loading	**.76**
Lowest loading	**.66**
Scale mean (1–7)	**4.91**

Reflective Flexibility	
.82 omega	
Alpha	**.81**
Mean loading	**.67**
Lowest loading	**.33**
Scale mean (1–7)	**4.89**

Model comparison

Model	CFI	TLI	RMSEA	SRMR
Four related scales (intended)	**.941**	**.933**	**.063**	**.055**
One overall dimension	.735	.709	.132	.109
Four scales, APS5–APS6 allowed to correlate	.946	.939	.060	.053

Conventional guides: CFI/TLI above about .90 acceptable and above .95 good; RMSEA and SRMR below about .08 acceptable and below .06 good. Scaled indices from a WLSMV estimator on polychoric correlations.

Verdict. Evidence supports interpreting the MAP-24 primarily as four related dimensions rather than a single undifferentiated trait. Fit is acceptable rather than excellent; the remaining misfit sits in a few small secondary relationships and the APS5–APS6 pair (page 3), none of which changes the structure.

How the four scales relate

	APS	FR	CSE	RF
APS	—	**.55**	**.59**	**.49**
FR		—	**.61**	**.51**
CSE			—	**.53**
RF				—

Latent correlations, corrected for measurement error. Values of .49–.61 mean the scales share a theme without collapsing into one another.

Why a high overall alpha is not enough

Alpha across all 24 items is .91, and omega under a single-factor model is .93. Both are high because the scales are positively related and there are many items — not because the instrument measures one thing. The single-factor model above is the direct test of that claim, and it fails it.

The overall 0–100 Adaptability Score

The transformation is straightforward: the 24-item mean rescaled so that 1 = 0 and 7 = 100 (sample mean 65.6, SD 14.1). As a high-level descriptive indicator it is usable and can be retained, clearly labelled as a summary rather than a diagnosis.

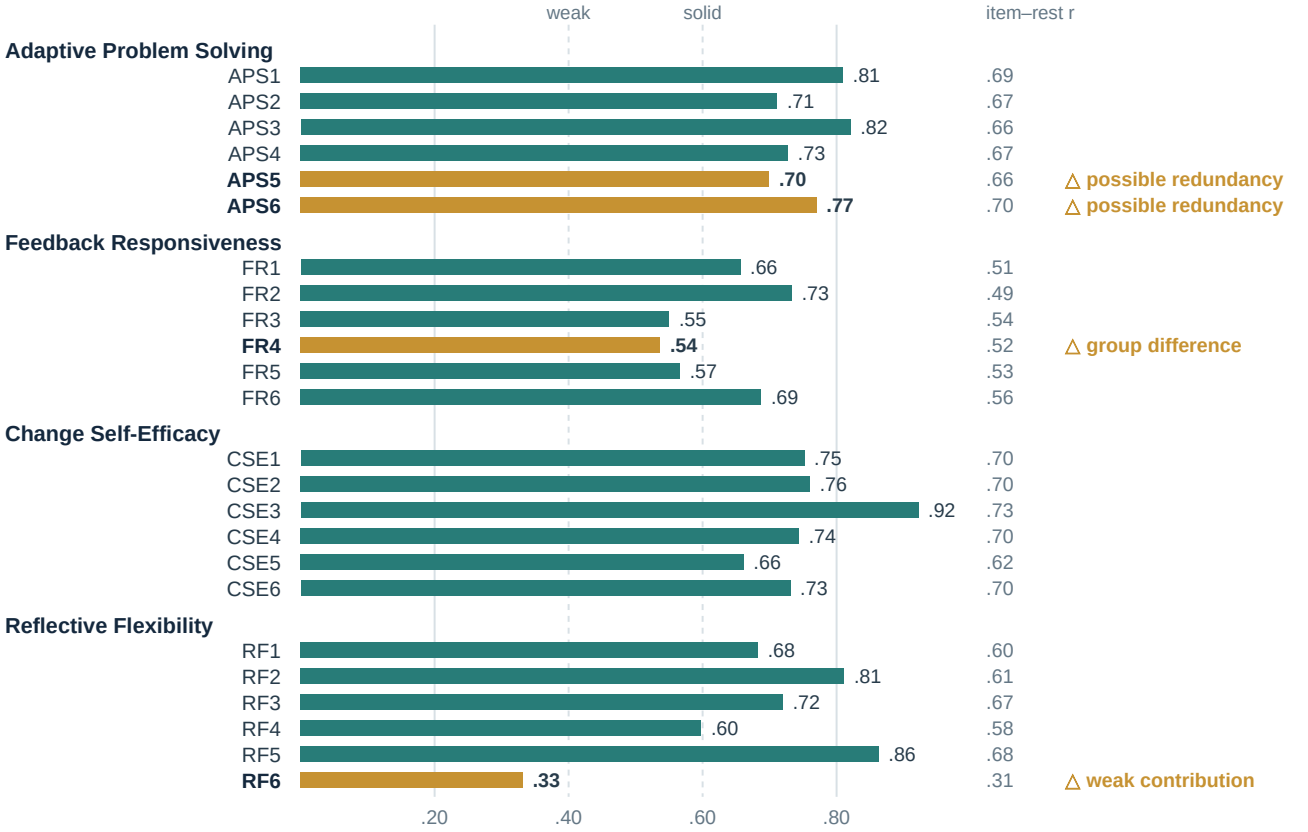
The cost is diagnostic. One number discards the shape of the profile: 70% of respondents differ by 20 points or more between their highest and lowest scale, and 25% of those scoring 60 or above on the total fall in the bottom quarter on at least one scale. For developmental feedback, that is the information a manager needs.

A score can be reliable without a particular cut-off being valid.

The 0–100 score is consistent (alpha .91, standard error of measurement about 4 points). That says nothing about whether 60 is a meaningful line. Whether scores below 60 identify people who need development is a separate empirical question about consequences, and no evidence for it currently exists (page 4). Decision-making should rest on the four profile scores; any threshold requires its own validation.

Item performance

Standardised loading of each item on its intended scale, with the item–rest correlation alongside. 23 of the 24 items load at .50 or above; four are flagged, three of them for reasons a loading alone would not reveal. No item shows floor effects; use of “strongly agree” ranges from 7% to 23% per item, typical for positively worded workplace scales.



RF6

Weak contribution

“I enjoy variety in my working week.”

Observation

Loading .33 and item–rest correlation .31, against .60–.86 for the other five RF items. Removing it would raise alpha from .81 to .84.

Why it matters

It contributes little information to the intended construct, and its wording asks about enjoying variety rather than reconsidering one’s thinking.

Recommendation

Review wording and conceptual relevance before the next data collection. Do not delete automatically: if the construct definition includes openness to varied demands, rewrite the item to target that; if not, replace it.

APS5 + APS6

Potential redundancy

“I find workable solutions when problems arise unexpectedly.” / “I come up with practical solutions when unexpected problems occur.”

Observation

Polychoric correlation .68 versus a median of .56 for other APS pairs; the largest residual in the whole model (APS5–APS6, modification index 40).

Why it matters

Two near-identical statements inflate reliability without adding coverage, and give this idea double weight in the scale score.

Recommendation

Keep one after content review, or rewrite the other to cover a distinct facet (for example, solving problems with limited resources). Expect alpha to fall slightly and validity to be unaffected.

FR4

Group-comparability concern

“I am comfortable being told that I have got something wrong.”

Observation

At the same underlying level of Feedback Responsiveness, Group B respondents endorse this item more strongly than Group A (uniform DIF, $\chi^2(1) = 20.2$, $p < .001$). The effect is small ($\Delta R^2 = .012$). No other item reaches the review threshold.

Why it matters

An item that works differently across groups can shift scale scores by a fraction of a point even when the scale as a whole is comparable.

Recommendation

Review whether “comfortable being told” is read differently across groups; retest after rewording. Not severe enough to invalidate the scale.

Loadings are from the four-scale ordinal CFA. Reference lines at .40 and .60 are conventions, not rules: a lower-loading item may be worth keeping for content coverage, and a very high loading may signal redundancy rather than quality.

Fairness, precision and validity

Do scores behave similarly across groups?

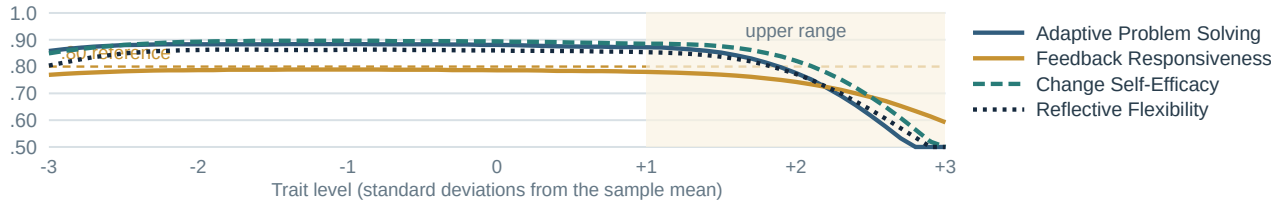
Invariance was tested across gender_group (Group A n = 255, Group B n = 225), constraining first the response thresholds and then the loadings; all 24 items were also screened individually.

Model	CFI	RMSEA	ΔCFI
Configural (same structure)	.941	.061	—
Equal thresholds	.935	.059	−.007
Equal thresholds and loadings	.940	.056	−.002
As above, FR4 thresholds freed	.943	.055	+.002

A CFI drop larger than about .01 is the usual signal of non-invariance; no step reaches it. Freeing FR4 improves fit significantly (χ^2 difference $p < .001$): that is where the localised difference sits, and FR4 is the only item flagged by the item-level screen. Latent mean differences are small ($|d| \leq 0.21$).

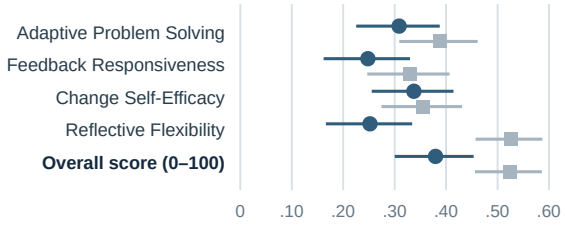
Scores were broadly comparable across the two demonstration groups at scale level, although one Feedback Responsiveness item showed a localised difference that should be investigated before broader deployment. Absence of detected differences here does not establish fairness in other populations or for other uses.

Where are scores most precise?



From a graded response model per scale; .80 is a common working standard. Three scales measure reliably (≥ .85) from well below average to about one standard deviation above it. Precision falls in the upper range, where roughly one respondent in seven sits (15%–14% across scales), so unusually high individual scores deserve more caution than mid-range ones. Feedback Responsiveness peaks at .79 and never reaches .80: its items separate respondents less sharply.

Does the assessment relate to something meaningful?



Circles: independent developmental effectiveness rating (1–7, mean 4.8). Squares: short openness measure, an adjacent construct. Bars: 95% confidence intervals.

All four scales relate positively to the effectiveness rating (r = .25–.34), strongest for Change Self-Efficacy. Reflective Flexibility correlates most with openness (.53) while the others sit at .33–.39: the pattern expected if the scales measure related but distinct things.

What this does and does not show. These are concurrent associations in one sample, from a rating that may share method variance with self-report. They are preliminary criterion-related evidence, not evidence that the MAP-24 predicts future performance, which needs outcome data collected later and a design that handles selection.

The 60/100 threshold

Overall score band	n	Mean effectiveness rating
0–50	62	3.89
50–60	96	4.48
60–70	128	4.77
70–80	116	5.11
80–100	78	5.56

The rating rises steadily with the overall score; there is no step at 60 (discontinuity test p = 0.87), and a score of 60 carries a measurement band of roughly ±9 points. Labelling the 33% of managers below 60 as “development needed” would be a policy choice with no validation behind it. Any threshold has to be set and tested against consequences in the population it will be used with.

Recommended next steps

1	**Review RF6 wording and content.** Decide whether it targets the construct as defined; rewrite or replace rather than simply delete.	Immediate
2	**Review APS5 and APS6 for duplication.** Retain one, or rewrite one to cover a distinct facet of problem solving.	Immediate
3	**Investigate and retest FR4 across relevant groups.** Check how “comfortable being told” is read; re-run the comparability analysis after rewording.	Before broader rollout
4	**Retain the four scale profiles as the primary interpretation.** Keep the 0–100 score as a labelled summary only; report it with its measurement band.	Immediate
5	**Do not use 60/100 as a pass/fail or risk threshold.** Any threshold needs its own validation against consequences, in the population it will be used with.	Immediate
6	**Collect longitudinal or outcome data if predictive claims are intended.** Ratings gathered later, ideally by a different source, are the minimum for a predictive claim.	Longer-term evidence programme
7	**Re-run validation after item revisions.** Changed items change the instrument; reliability, structure and comparability need to be re-established on new data.	Before broader rollout

How the assessment was reviewed

Sample	480 synthetic respondents; Group A 255 / Group B 225; three management levels and three age bands. No missing responses; no impossible values; no respondent answered every item identically. No reverse-keyed items.
Classical analysis	Item means and spread, floor and ceiling use, item–rest correlations, coefficient alpha and alpha-if-dropped for each scale and for the 24-item total.
Structure	Confirmatory factor analysis on polychoric correlations (WLSMV, lavaan): intended four-scale model, a single-factor model, and a diagnostic model allowing the APS5–APS6 residual to correlate. Omega computed from the ordinal model.
Comparability	Multi-group ordinal CFA across gender_group (configural, equal thresholds, equal thresholds and loadings, partial). The two lowest response categories (2.5% of all responses) were merged for the multi-group models only. Item-level DIF screened with ordinal logistic regression on the scale rest score.
Precision	Graded response model per scale (mirt); conditional reliability from test information.
Validity	Correlations with a concurrent effectiveness rating and a short openness measure, with 95% confidence intervals.
Score interpretation	Distribution of the 0–100 score, its standard error of measurement, profile spread, and a test for any discontinuity in the criterion at 60.

Omega: the share of scale-score variance attributable to the construct, estimated from the factor model. **Factor loading**: how strongly an item reflects its scale, from 0 to 1. **CFI/TLI**: how much better the model fits than a model with no structure, with 1 the ceiling. **RMSEA/SRMR**: average misfit per parameter or per correlation, where lower is better. **DIF**: an item working differently for people at the same underlying level in different groups.

This demonstration illustrates the type of evidence that can be generated from a psychometric health check. Validation is an accumulating body of evidence rather than a one-off certification. Conclusions depend on the population, intended use and consequences of score interpretation.

PREPARED BY

Dr Gregory Gorman

Psychometrician & Measurement Consultant
PhD in Psychometrics
Scale development · IRT · CTT · CFA/SEM · Measurement invariance · Validation
greg@example.com