Advanced Processing Test Technical Report

Created and analyzed by u/GuessSoButNo

8/28/2025

Contents

Overview

An analysis of the APT was conducted in order to validate the test. With data from 1,197 testees answering 40 questions across five different subtests (Analogies, Number Series, Vocabulary, Arithmetic, and Matrix Reasoning), some interesting patterns were found. The test shows solid reliability (consistency) and has a strong general intelligence factor. Confirmatory Factor Analysis found that approximately 74% of a testee's overall score comes from their general intelligence, with the rest likely coming from specific verbal or math skills. The math and number-based sections showed the strongest connection to overall intelligence, while surprisingly, the Matrix Reasoning section was the weakest. Regardless, the APT appears to be a reasonable 20-minute IQ test that measures both general intelligence and specific cognitive abilities.

1 Background and Data

To reiterate, the APT consists of five subtests that test different mental abilities:

- **Analogies (AG)**
- **Number Series (NS)**
- **Vocabulary (VC)**
- **Arithmetic Reasoning (AR)**
- **Matrix Reasoning (MR)**

Note: AG and VC are expected to load on verbal comprehension (Gc), whereas AR, MR, and NS are expected to load on fluid intelligence (Gf).

Responses from 1,197 people who took the test online were analyzed. Originally, there were over 4,000 test attempts, but after removing duplicate attempts, incomplete tests, and non-native English speakers, we end up with our final sample of N = 1,197.

The average standard score was 120.01, with the standard deviation being 12.08. The test by default has a mean of 100, with a standard deviation of 15. The test also originally consisted of 80 items, taken in 40 minutes. This analysis is for the short form, that is, the one provided here.

2 Analysis

For this particular application, tetrachoric correlations were computed based on binary data. Several reliability measures were then computed. Afterwards, both exploratory and confirmatory factor analysis were computed.

3 Results

3.1 Subtest Relationships

First, an analysis of how different subtests relate to each other was investigated. Here are the results:

- Number Series and Arithmetic had the strongest relationship (.596), which makes sense as they both involve computations.
- Analogies and Vocabulary were also connected well (.509), which makes sense, as both tap into the verbal domain.
- Matrix Reasoning did not seem to have much of a strong connection to the other subtests. Although this may be surprising, it does make sense given the limited number of items, and it being a comparatively distant factor due to exposure.

Table 1: Subtest intercorrelation matrix

	AG	NS	VC	AR	MR
AG	1.000	0.382	0.509	0.419	0.270
NS	0.382	1.000	0.341	0.596	0.383
VC	0.509	0.341	1.000	0.313	0.245
AR	0.419	0.596	0.313	1.000	0.335
MR	0.270	0.383	0.245	0.335	1.000

3.2 Test Reliability

The test showed excellent reliability, with ordinal $\alpha = 0.911$.

3.3 The General Intelligence Factor

It was found that general intelligence accounts for 74% of test performance (item bifactor CFA) and up to 79% with an alternative computation. The remaining variance likely comes from specific verbal or mathematical abilities not measured.

So of the overall score: $\approx$ 74% ***g*,** $\approx$ 17% **other reliable variance,** $\approx$ 9% **random error** *(in this sample for the bifactor method).*

Note (i) *Standardized loadings* are parameters for items or subtests on g. (ii) So it is more technically accurate to emphasize the usage of the following terms instead: *general–total correlation* $r_{gT} = \sqrt{\omega_h}$ which is the correlation between latent g and the unit weighted total score. (iii) The *general factor saturation* of the total score, ω_h, is the proportion of total-score variance attributable to g. Essentially, it is more accurate to use "loading" for parameters and use r_{gT} and ω_h when describing the total score.

Breaking it down by subtest, we can see how much each one relates to g:

- **Number Series**: Highest connection to g (.604)
- **Arithmetic**: High (.573)
- **Analogies**: Moderate (.337)
- **Vocabulary**: Moderate (.331)
- **Matrix Reasoning**: Low (.244)

Table 2: General factor from item bifactor CFA

	ω_h (general factor saturation)	General–total correlation ($r_{gT} = \sqrt{\omega_h}$)
Item-level bifactor CFA	0.743	0.862

3.4 Model Fit

The item-level bifactor CFA fit well (SRMR is slightly high but expected in this context): $n_{\text{par}} = 120$, $\chi^2(700) = 1503.36$, $p < .001$, CFI = .954, TLI = .949, RMSEA = .031, SRMR = .089.

Table 3: Model Fit Comparison

Level	Model	χ^2 (df)	CFI	TLI	RMSEA	SRMR
(A) Subtest level						
Subtest	Higher-order (Gc, Gf on g)	13.90 (3)	.992	.975	.055	.014
(B) Item level (items)						
Item	Bifactor (g + Gf + Gc)	1503.36 (700)	.954	.949	.031	.089

Note: The item level estimators used are not the same.

4 What Does This Mean?

The APT appears to successfully measure both general intelligence and specific domains. The performance from Number Series and Arithmetic suggests these are particularly g-loaded subtests.

The weak performance of Matrix Reasoning is surprising, but is explainable. It has the fewest items, and it is the least related comparatively to other subtests. For example, AR and NS will naturally have a stronger relationship with each other than MR due to most verbally encoded computations, whereas MR involves a more distant visual modality. Moreover, the first item has poor discrimination, with p ≈ .99, which is credited due to the ceiling effect. The same is not attributable to other subtests, which have more items, and items with better discrimination. Another plausible, though not fully substantiated theory, may also be that the sample also consists of many users who are quite familiar with the subtest. Note, however, that dropping the weakest MR item left the g-factor saturation essentially unchanged, meaning that it is likely the total-score g-saturation isn't contingent on a single low-discrimination item (the same case applies to other lower performing items).

5 Summary

The APT is a good test, considering it is only 40 questions meant to be taken in 20 minutes.

That said, there may be room for improvement. The Matrix Reasoning subtest may need some items reworked, with harder and more discriminatory items or perhaps an entirely different nonverbal measure. Granted, there is only so much one can do, considering there are really only six informative items (the first MR item being too easy to provide any information). Additionally, expanding the test with more MR items could also help.

For testees, the results suggest that your overall score genuinely reflects a combination of your general cognitive ability.

A Additional Information

A.1 Data Summary

From around 4,000 raw attempts, duplicate attempts were removed, incomplete sessions, non–English speakers, and low-effort cases, providing us with the analyzed sample of $N = 1197$.

A.2 Subtest Descriptives

Table 4: Descriptives by subtest (N = 1197)

Subtest	Mean	SD	Skew	Kurtosis
Analogies (AG)	4.330	1.840	0.347	−0.394
Number Series (NS)	4.400	1.260	−0.178	−0.334
Vocabulary (VC)	5.230	1.460	−0.278	−0.179
Arithmetic Reasoning (AR)	5.060	1.480	−0.695	1.030
Matrix Reasoning (MR)	4.150	1.750	0.003	−0.540

Note: Items per subtest are: AG = 8, NS = 8, VC = 8, AR = 9, MR = 7.

A.3 Intercorrelations, Reliabilities, and Loadings

Table 5: Subtest intercorrelations

	AG	NS	VC	AR	MR
AG	1.000	0.382	0.509	0.419	0.270
NS	0.382	1.000	0.341	0.596	0.383
VC	0.509	0.341	1.000	0.313	0.245
AR	0.419	0.596	0.313	1.000	0.335
MR	0.270	0.383	0.245	0.335	1.000

A.4 General Factor Indices

Table 6: General factor results

(a) Reliability indices		
α (KR-20)	0.826	lower for dichotomous data in this context
Ordinal α	0.911	
ω_t (SL EFA)	0.921	another way to represent reliability
$\sqrt{\omega_h}$ (SL EFA)	0.777	general–total correlation (r_{gT})
ω_h (Monte Carlo)	0.711	general-factor saturation (estimation)
$\sqrt{\omega_h}$ (Monte Carlo)	0.843	general–total correlation (estimation) (r_{gT})
ω_h (bifactor CFA, alternative analytic method)	0.793	general-factor saturation
$\sqrt{\omega_h}$ (alternative CFA)	0.890	general–total correlation (r_{gT})
ω_h (bifactor CFA; main one used here)	0.743	general-factor saturation
$\sqrt{\omega_h}$ (bifactor CFA)	0.862	general–total correlation (r_{gT})
(b) Mean item loadings on g (bifactor CFA)		
AG	0.337	
NS	0.604	
VC	0.331	
AR	0.573	
MR	0.244	
(c) Correlation of g factor scores with subtest totals		
AG	0.572	
NS	0.851	
VC	0.476	
AR	0.860	
MR	0.503	

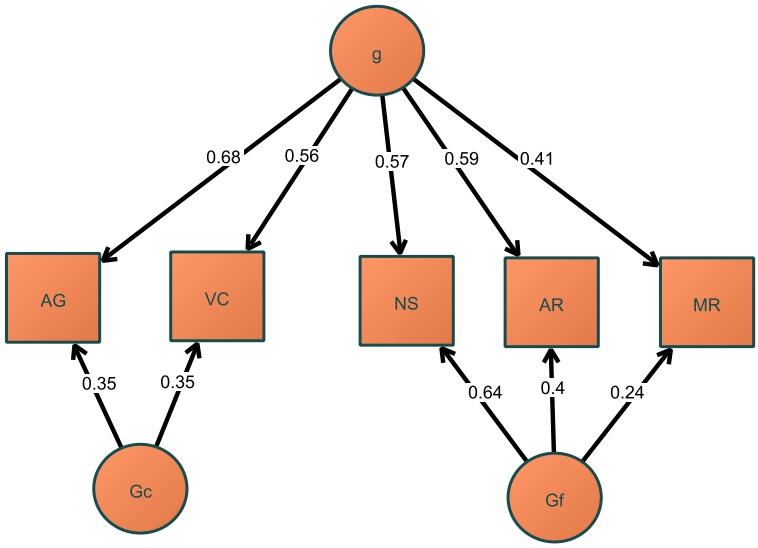

Figure 1: Bifactor CFA.

A.5 Variance Decomposition of Total Score

Table 7: Decomposition of total score variance

Component	Proportion
General factor g	0.743
Specific/group factors ($\alpha - \omega_h$)	0.168
Random error ($1 - \alpha$)	0.089

Note: This uses the calculated ordinal $\alpha = 0.911$ and CFA $\omega_h = 0.743$.

A.6 Higher-Order CFA (Subtests: Gc & Gf on g)

Table 8: Higher-order CFA (Gc,Gf on g): fit and standardized loadings

(a) Fit indices	
Parameters (n_{par})	17
χ^2 (df), p	13.9 (3), $p = .003$
CFI / TLI	0.992 / 0.975
RMSEA [90% CI]	0.055 [0.028, 0.086]
SRMR	0.014
(b) Standardized $g \rightarrow$ subtest loadings	
AG	0.641
NS	0.636
VC	0.532
AR	0.619
MR	0.389

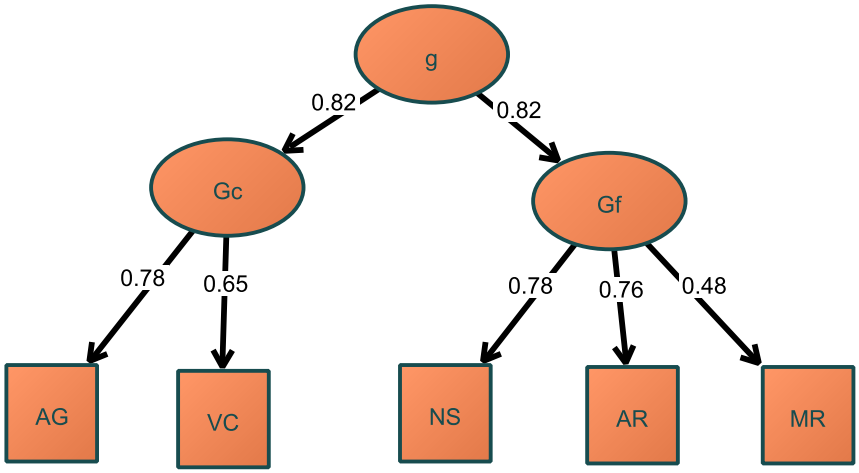

Figure 2: Higher-order CFA.

A.7 Schmid–Leiman EFA

Table 9: Schmid–Leiman EFA:

(a) Indices	
ω_t	0.921
ω_h	0.603
$\sqrt{\omega_h}$	0.777
(b) Mean item g-loadings by subtest	
AG	0.325
NS	0.468
VC	0.356
AR	0.440
MR	0.246

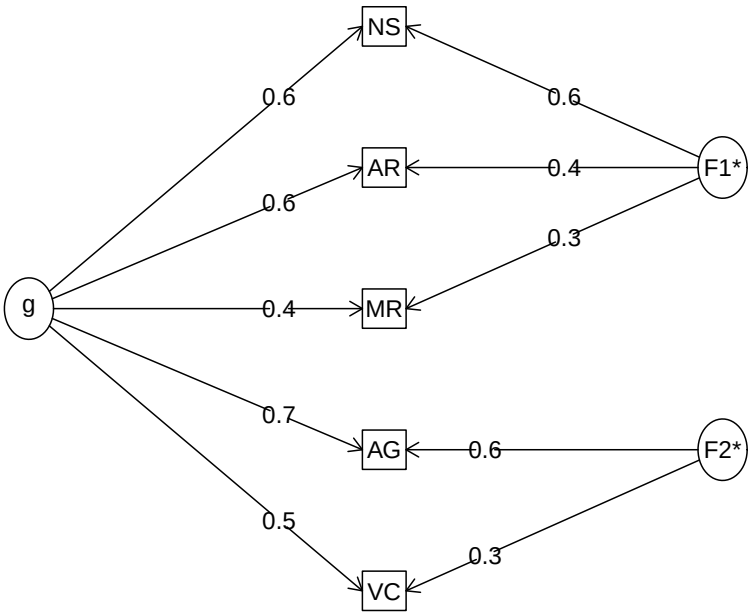

Figure 3: Schmid-Leiman diagram summarizing general and group factor loadings. Note, however, that to interpret this with caution, as only two group factors were used. This isn't recommended as equal loadings are set, but it is a nice addition to see what it would report. Higher group factors specified returned similar results.

A.8 Classifying Arithmetic Reasoning (AR) within Gf

AR was chosen to go with fluid intelligence (Gf) instead of verbal comprehension (Gc) because it was more empirically aligned with quantitative reasoning. For example, from the intercorrelation matrix, AR shows its strongest link with Number Series (NS; r = .596), and more moderate links to verbal subtests (VC r = .313; AG r = .419) and MR (r = .335). In the bifactor CFA, subtest loadings on g were NS = .604, AR = .573, AG = .337, VC = .331, MR = .244, placing AR essentially next to NS on the latent g continuum.

AR items also typically require quantitative reasoning and working memory (Gf/Gwm), but in this particular case, Gwm was mediated by the fact you can use paper. Additionally, the arithmetic items here did not need to be read out, unlike some arithmetic items on professional batteries. This would consequently reduce the emphasis on Gwm and Gc (as one needs to both retain the spoken item and understand it, understanding it will help one naturally memorize it better).

A.9 Country Level Descriptives

Table 10: Mean standard scores by country

Country	Mean	SD
Singapore	124.670	12.030
New Zealand	120.460	11.750
Australia	120.320	11.620
United States	120.180	12.110
United Kingdom	120.110	12.460
Canada	118.210	12.170

Note: Descriptive only from a filtered online sample.

Figure 4: Country mean standard scores with 95% CIs (exploratory; non-normative).

A.10 Score Bounds (SEM)

Given the observed SD = 12.08 and ordinal $\alpha = 0.911$, the standard error of measurement can be computed simply as SEM = $\text{SD}\sqrt{1-\alpha} \approx 12.08\sqrt{0.089} \approx 3.60$ SS points, implying a 95% score band of $\pm 1.96 \times \text{SEM} \approx \pm 7.06$ SS points around an obtained score for the sample.

A.11 Completion Time (Speededness)

Table 11: Completion time summary (in minutes)

N	Mean	SD	Median	5th %ile	95th %ile	% at cap
1197	19.200	1.670	20	15.400	20.000	59.2%

Note: Cap = 20 minutes; time is in minutes.

Table 12: Association between completion time and standard scores

Metric	Estimate	95% CI
Pearson r (all)	0.007	[-0.050, 0.063]
Spearman ρ (all)	−0.097	—

A.12 Reasoning Speed vs. Processing Speed (PSI)

APT imposes a 20 minute limit, but the test is likely not unreasonably dependent on PSI. For example, time to score correlation is essentially zero overall ($r = .007$). Second, excluding capped administrations, longer completion time predicts higher scores ($r = .207$; Spearman $\rho = .198$), and the 15 to 20 minute band shows the highest mean standard score (123), while $\leq$ 15 minutes is lowest (115). Additionally, capped testees actually score lower than testees who do not hit the time limit. These findings themselves don't seem to be concordant with a processing speed factor, and are instead probably more consistent with reasoning.

A.13 Further Processing Speed Analysis

To further substantiate aforementioned claims, a regression of latent g on time was applied. The result was small and likely insignificant. The standardized effect per +1 SD of time was $\beta = 0.012$. A processing speed loading would typically predict a negative result, but the observed effect is near zero and slightly positive in this particular case. This, in addition to the data above which shows no association between time and standard scores in the full sample (Pearson $r = .007$) and a positive association when excluding capped administrations (Pearson $r = .207$). This probably means the APT isn't speed loaded, and instead taking a bit longer is weakly associated with higher latent ability.

A.14 Subtest Reliabilities

Table 13: Subtest reliabilities (KR–20 and ordinal α)

Subtest	KR-20	Ordinal α
AG	0.474	0.681
NS	0.662	0.848
VC	0.591	0.795
AR	0.667	0.834
MR	0.363	0.542

A.15 Item Diagnostics

Table 14: Median item difficulty (p) and median point–biserial by subtest

Subtest	Median p	Median point-biserial
AG	0.759	0.286
NS	0.576	0.381
VC	0.849	0.275
AR	0.384	0.351
MR[†]	0.691	0.183

Note: p = proportion correct (higher = easier). Point–biserial is corrected.

[†]The first MR item was flagged for very low discrimination (point–biserial < .10).

Table 15: Difficulty coverage by subtest (proportion of items in each band)

Subtest	% with $p < .30$ or $p > .80$	% with $.30 \leq p \leq .80$
AG	50.000	50.000
VC	87.500	12.500
NS	75.000	25.000
AR	66.700	33.300
MR	57.100	42.900

Note: Extreme p indicates that there is potential ceiling/floor; the middle band is most optimal in providing information.

A.16 Unidimensionality

Bifactor indices show unidimensionality for scoring: $\omega_h = 0.743$, ECV = 0.641, and PUC = 0.492. A one–factor fit to the tetrachoric inter–item matrix also yielded a first residual eigenvalue of 2.40, which is by all means, only modestly above a common heuristic of ≈ 2.0 and is expected since the test spans verbal and nonverbal content. So, these results support a reasonable general factor with expected domain residuals. Another important remark is with PUC = 0.492. The value means that many items (roughly half) have a shared domain factor. As such, one should interpret subtest results warily. The total score remains fine, however.

A.17 IRT

A 2PL validation showed peak test information near $\theta \approx 0.60$, with $I(0) = 6.16$. This implies $\mathrm{SEM}(0) = 1/\sqrt{6.16} \approx 0.40$ and conditional reliability at $\theta = 0$ of $I/(I+1) \approx 0.86$. Thus, the test is converged slightly above the population mean, so it is consistent with the sample's elevated mean.

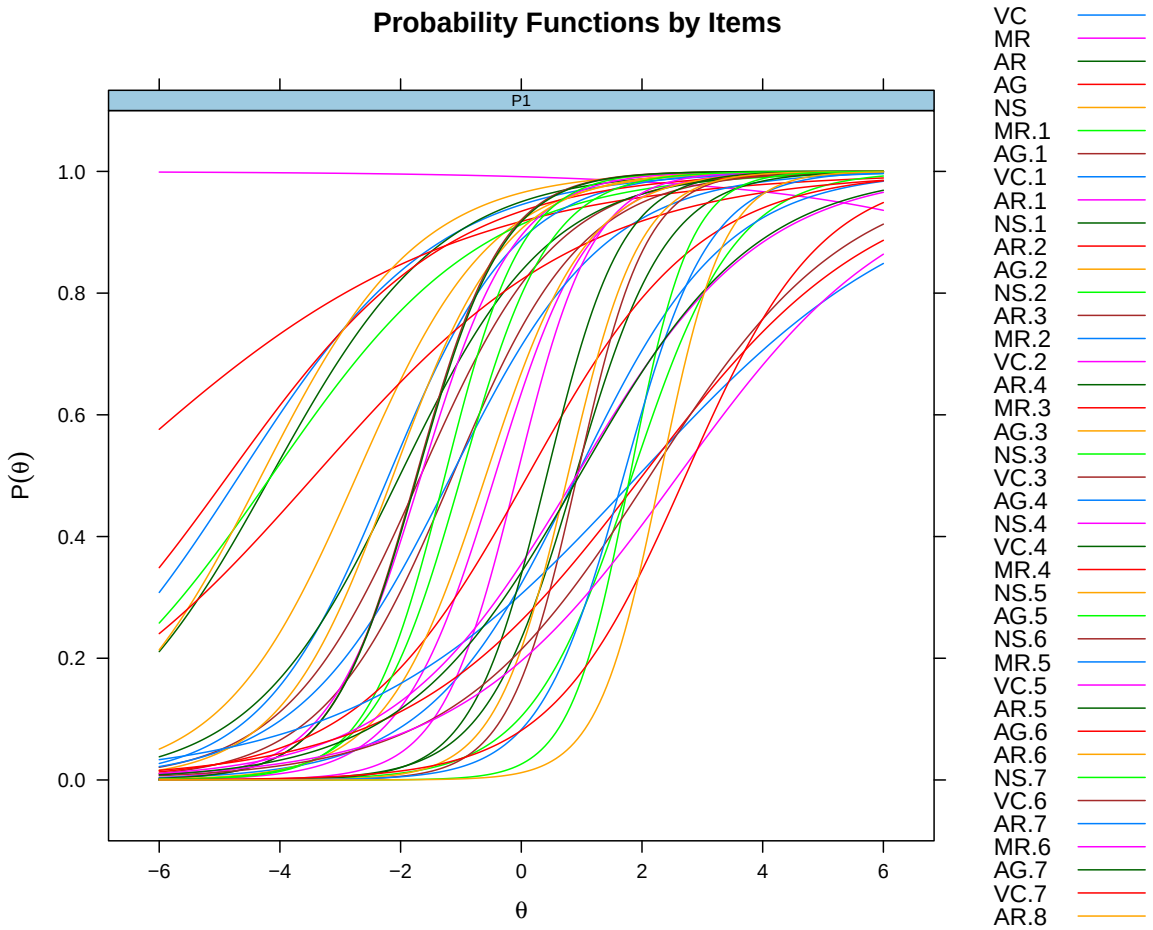

Figure 5: 2PL probability functions. Take note that the item with the weakest discrimination, MR, is quite apparent.

Split-sample stability. Random split-half replication yielded a correlation of item g–loadings $r = .854$ across halves, with $\omega_h = .783$ (half A) and $\omega_h = .818$ (half B).

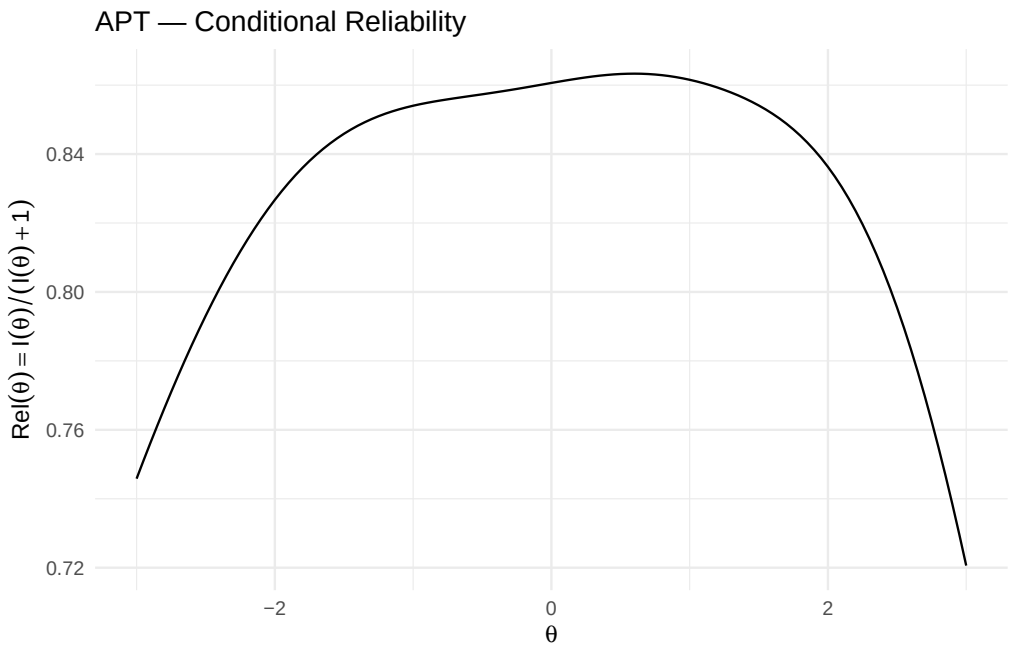

Figure 6: Conditional reliability Rel(θ) was computed from (Rel(θ) = $I(\theta)/(I(\theta)+1)$). Precision maxes out near $\theta \approx +0.6$ (SS $\approx$ 127.5 using an intercept 119.85 and a slope of 12.78, with Rel(θ)$\geq$ .85 for roughly $-1.5 \lesssim \theta \lesssim +2.0$. At $\theta = 0$, $I(0) = 6.16$ which implies Rel(0) $\approx$ .86 and a SEM(0) $\approx 1/\sqrt{I(0)} \approx 0.40$ on the θ scale ($\approx$ 5.15 SS points via the link)

.

A.18 Factor score quality

Factor determinacy for the bifactor g was high (FD = 0.97, FD^2 = 0.94), so there is proper correspondence between the estimated and true scores. As mentioned above, random halves showed good stability of item g–loadings (r = .854) and similar factor saturation (ω_h = .783 vs. .818). Additionally, one can also coin (a new term, as it may be useful in future analysis) the *score determinacy index* (SDI), defined here as FD^2; in this sample SDI = 0.94.

Alignment of g with the APT *standard score* was r = .939. This association was unchanged when controlling for time (standardized β_g = .939; partial r = .939), which suggests this is consistent with likely negligible speed effects on g.

Table 16: Additional validity indices

Check	Result
IRT information peak	$\theta \approx 0.60$
Information / SEM / Rel at $\theta = 0$	$I(0) = 6.16;\ \mathrm{SEM}(0) \approx 0.40;\ \mathrm{Rel}(0) \approx 0.86$
Split-half r (item g–loadings)	$r = .854$
ω_h by split	0.783 / 0.818
$g \sim$ time	$\beta = 0.012,\ p = .733$ (std)

Note: Alignment of g with the APT *standard score* was $r = .939$, unchanged when controlling for completion time (standardized $\beta_g = .939$; partial $r = .939$). That is, the bifactor g score and the APT total are closely related ($r = .939$). The APT also correlates highly with other tests.

A.19 Range Restriction and SLODR

The observed general factor saturation from the bifactor CFA was $\omega_h = 0.743$ (so $r_{gT} = \sqrt{\omega_h} = 0.862$), where r_{gT} is the correlation between the general factor (g) and the unit-weighted total score.

Range restriction (RR) Because the standard-score SD in our sample was 12.08 versus the nominal 15 (restriction index $u = 0.805$), one can apply the Thorndike Case II direct range-restriction correction to r_{gT}:

$$r_{gT}^{RR} = \frac{U\, r_{gT}}{\sqrt{1 + r_{gT}^2\,(U^2 - 1)}}, \quad U = 1/u.$$

This yields $r_{gT}^{RR} = 0.904$ and $\omega_h^{RR} = (r_{gT}^{RR})^2 = 0.817$. These values will also increase for the other aforementioned ω_h.

SLODR projection To approximate the effect of ability distribution (SLODR), a regression r_{gT} on external ability could be applied. Projecting from the sample mean ability here might actually result in a marginally deflated loading, which is contrary to conventional expectations. However, given that the APT was designed with the idea of higher ability people in mind, such results may make sense. In this case, the projection resulted in a general total correlation $r_{gT} = 0.83$ of 0.83 (uncorrected). Although this direction is the reverse of what is commonly expected, the effect is modest ($\Delta r \approx -0.035$) and has a reasonable interpretation here: (i) the APT item pool is targeted above average ability, so higher-ability testees contribute less guessing, (ii) nearer the population mean there is greater profile heterogeneity, which may increase specific variance relative to g. These factors can therefore actually yield a small "reverse-SLODR" for this dataset, but it certainly does not invalidate the substantive conclusion that the total score remains a strong index of general ability, as shown in previous sections. More interestingly, however, is that if one additionally corrects the SLODR-projected r_{gT} for range restriction $u = 0.805$), one will come to find that the estimate is $r_{gT}^{\text{SLODR+RR}} \approx 0.879$ ($\omega_h \approx 0.773$), which is similar as the raw sample value. The SLODR projection itself (removing the sample's +1.33 SD mean advantage) lowers r_{gT} by ≈ 0.035, while the Thorndike correction for SD restriction ($u = 0.805$) raises it by a similar amount, yielding $r_{gT}^{\text{SLODR+RR}} \approx 0.879$ ($\omega_h \approx 0.773$). Thus the opposing selection is largely offset, indicating the observed ω_h and g is robust to plausible population adjustments in this analysis. Finally, rebalancing very easy and/or very hard items around $\theta \approx 0$ would likely flatten the r_{gT}–by–ability slope.

Caveats. (1) Thorndike Case II assumes linearity, homoscedasticity, normality, and selection directly on the observed score. (2) The SLODR projection depends on the external ability linkage and the estimated slope.

A.20 Final general–total correlation and SLODR

Plausible bounds for the overall general–total correlation Using two standard estimators from the bifactor CFA, one can obtain $\omega_h = 0.743$ ($r_{gT} = 0.862$) and $\omega_h = 0.793$ ($r_{gT} = 0.891$). Applying Thorndike Case II range–restriction (RR; $u = 0.805$) to each gives $r_{gT}^{RR} = 0.904$ ($\omega_h^{RR} = 0.817$) and $r_{gT}^{RR} = 0.925$ ($\omega_h^{RR} = 0.855$). Thus a practical bracket is $r_{gT} \in [0.862,\ 0.891]$ (RR–only: $[0.904,\ 0.925]$).

An important note on SLODR. Credible tests of Spearman's Law of Diminishing Returns (SLODR) require specific designs with stringent tests of measurement invariance across ability strata. Simple regressions of r_{gT} (or related indices) on ability which was done here are *simple and experimental* and could plausibly be confounded by range restriction, scale nonlinearity, and other factors. Moreover, it is still debatable as a whole whether traditional SLODR heuristics can reproduce SLODR-esque patterns from subtest characteristics even when the underlying structure is invariant. Consequently, these methods may not reliably separate artefacts from genuine SLODR effects. Following this, the SLODR computation performed here should be taken with a grain of salt and not as entirely concrete. That is, there is a reasonable chance it could be inaccurate.

B Final Remarks

The APT has good psychometric performance considering it is only a 20-minute test with 40 questions. A dominant general factor is supported and validated through quite a few considerable methods (bifactor CFA $\omega_h = .743$; ECV = .641; PUC = .492), and reliability is high (ordinal $\alpha = .911$). Precision is strongest from roughly $-1.5 \lesssim \theta \lesssim +2.0$ with a peak near $\theta \approx +0.6$, and remains solid at the population mean ($\text{Rel}(0) \approx .86$). A regression was also conducted which found that completion time does not depress g (std. $\beta = 0.012$, $p = .733$), so it stands as sufficient evidence against a significant, confounding speed factor. Selection corrections also interestingly indicate robustness, that is, range restriction increases r_{gT}, while the SLODR projection (to population mean ability) slightly lowers it; together they offset each other. Matrix Reasoning is the main opportunity for improvement (shorter subtest, one low-discrimination item, and potentially other factors). Finally, I think it is important to note that although this test is psychometrically sound, it of course may not be reasonably accurate enough to provide sufficient information about one's cognitive abilities nor be used in any diagnostic manner. In such cases, a professional evaluation is superior. Moreover, the score one receives here, should also not be used to define one's self-worth.