SAT Math Advanced Rendition Test: Technical Report

u/soapyarm

February 17, 2024

I. Introduction

The SAT Math: Advanced Rendition Test (SMART) is designed to be an emulation of the 1974 - 1994 SAT math section with an extended ceiling.

Like the SAT-M, this test serves as a comprehensive assessment of quantitative reasoning skills, targeting testees with a minimum of high school-level math proficiency. While the overwhelming majority of SMART items are original, they were meticulously designed to reflect the format, style, and scope of the original SAT-M. However, SMART distinguishes itself by featuring questions that are, on average, more difficult than those found in the SAT-M. While the original SAT-M has a ceiling of **152 IQ**, SMART has an estimated ceiling of around **168 IQ**. The main reason its ceiling doesn't reach as high as expected is due to its more generous time limit.

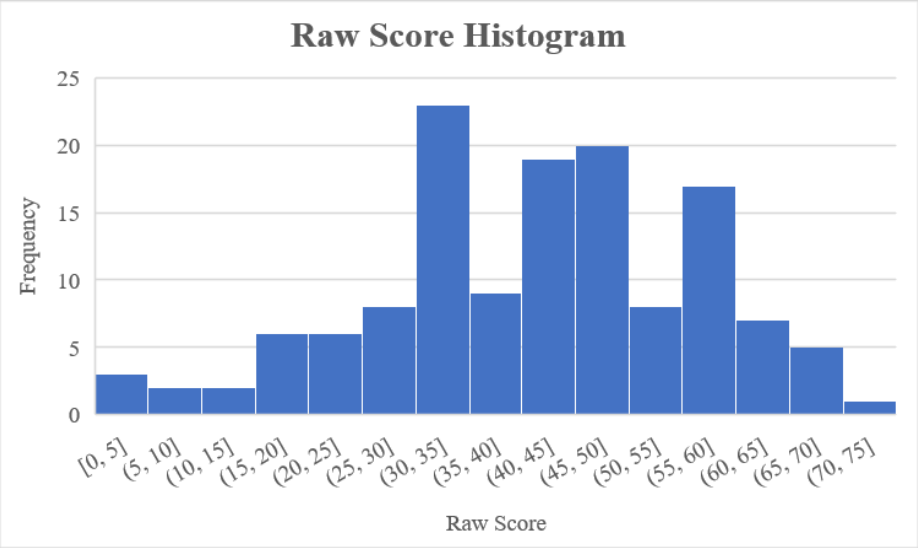
This report will detail the measures of SMART's reliability, its correlations with other established tests, and its estimated *g*-loading. The findings suggest that SMART effectively replicates the original SAT-M and serves as an excellent tool for assessing quantitative reasoning abilities.

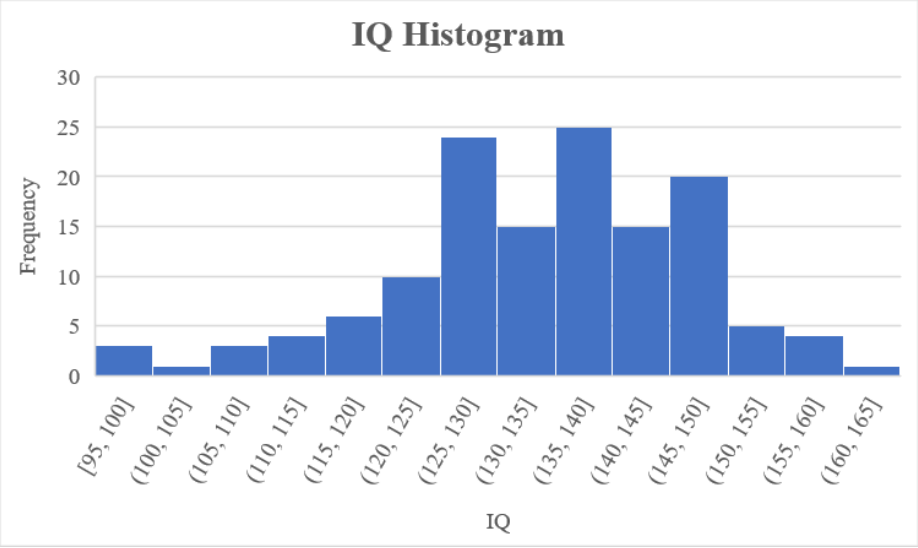
II. Collected Data

As of January 18th, 2024, **n = 135** valid first attempts were collected. Throwaway attempts, incomplete attempts, and potentially cheated attempts were detected and removed.

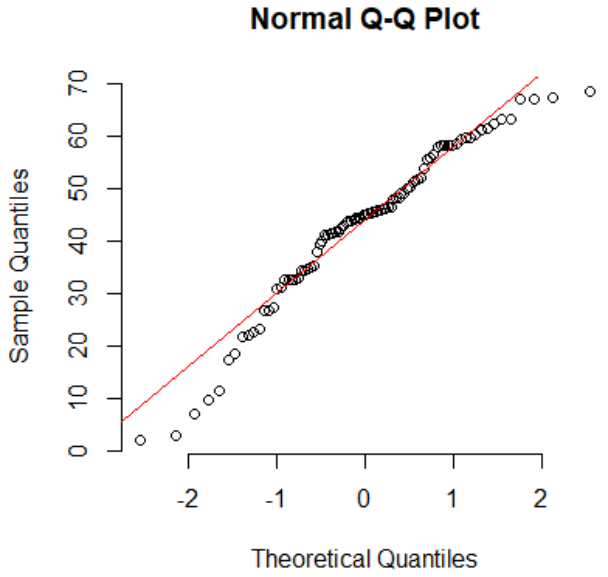
Raw				
Min	Max	Mean	Median	St. Dev.
2	74	41.6	44	15.2

IQ				
Min	Max	Mean	Median	St. Dev.
95	165	134.5	137	12.7





A **Q-Q plot** allows us to determine if the data follows a normal distribution:



The data points at the extremes do not align closely with the red line, suggesting that these scores deviate significantly from normality. This deviation is more pronounced at the very low ranges. It follows that both extremely high and extremely low scores occur less frequently than expected in a normal distribution.

Thus, the test was normed using **equipercentile equating**, applying it to all available professional test scores within the reliable range (115 - 160 IQ) and then extending it speculatively to the rest of the range. The updated norms are available here.

III. Reliability Measures

In assessing SMART's reliability, three statistical tests were employed: Cronbach's α, McDonald's ω, and Split-Half Reliability. The former two are measures of internal consistency, indicating how closely related the items are. The latter estimates the consistency of test results by correlating two equivalent halves of the test. One half consisted of odd-numbered questions, and the other half consisted of even-numbered questions.

The values below suggest **excellent** reliability, suggesting that the items form a highly cohesive set that measures a single underlying construct. In layman terms, the reliability represents the true score variance. Here, **~93%** of the variance in scores is reliable; the rest is random error.

Cronbach's α	**McDonald's ω**	**Split-Half Reliability**
0.928	0.927	0.911

IV. CORRELATIONS

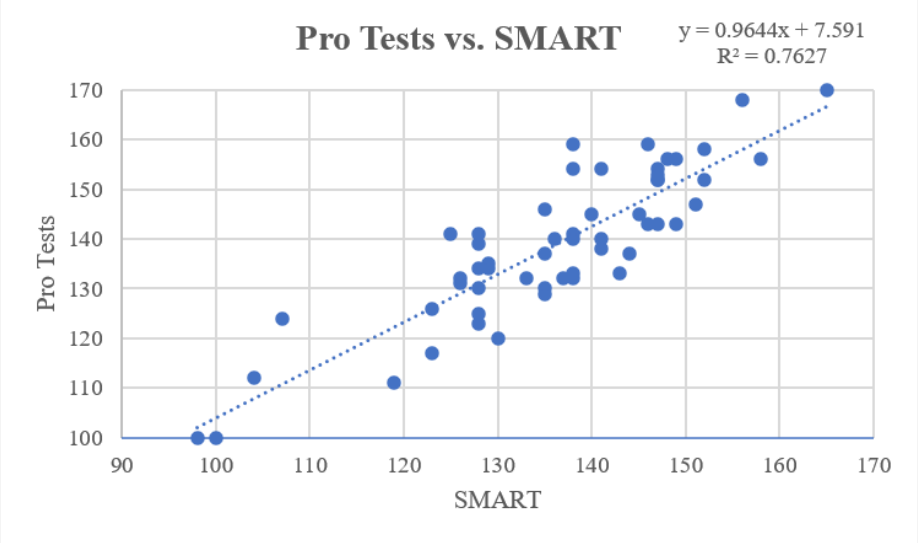
In assessing SMART's validity, its scores were compared with those from established professional quantitative tests.

The average IQ from professional tests was calculated under the following conditions:
(1) If any scores from SAT-M, GRE-Q, or QAT were present, only they were considered.
(2) If not, quantitative indices from full-scale tests were considered.
(3) If the score exceeded the ceiling of SAT-M and GRE-Q, only the score from QAT was considered (if present).
(4) If no quantitative scores were present, fluid or full-scale IQ scores were considered.

(1) aligns with SMART's design objective, which is to emulate the SAT-M. Therefore, the SAT-M, GRE-Q, and QAT, all of which closely resemble the SAT-M, are the most appropriate for comparison. (2) reflects SMART's nature as a quantitative test, making it logical to compare it with other quantitative tests. (3) accounts for ceiling effects that may arise with the SAT-M and GRE-Q. (4) acknowledges that quantitative reasoning is a subset of fluid reasoning, which is in turn a subset of general intelligence (*g*). Hence, when quantitative scores are unavailable, it is reasonable to consider fluid or FSIQ scores for comparison.

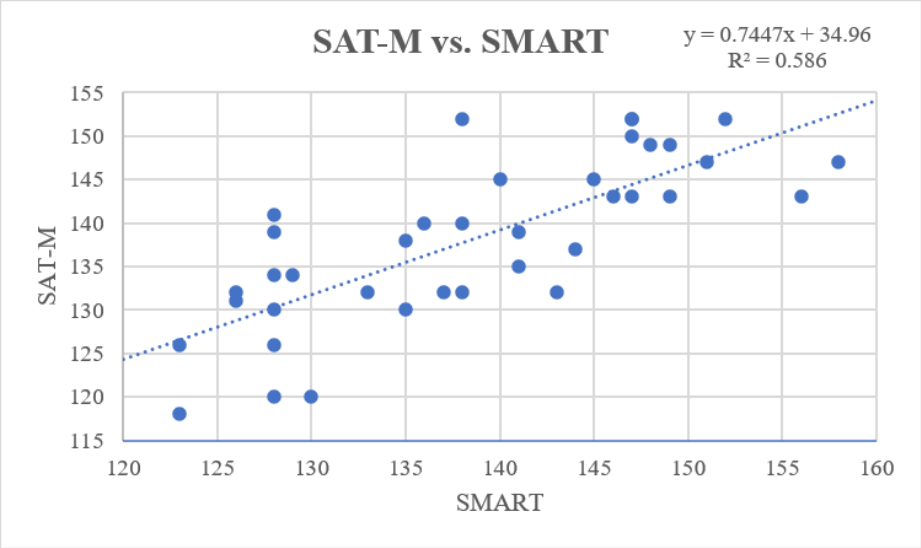
n = 55 average professional test IQ scores were collected from the following tests: SAT-M, GRE-Q, QAT, SB-V, WAIS-IV, WISC-V, RAIT, Raven's 2, and JCTI. SMART correlated with professional tests at **r = 0.873**.

Pro Tests				
Min	Max	Mean	Median	St. Dev.
100	170	138.8	140	15.0



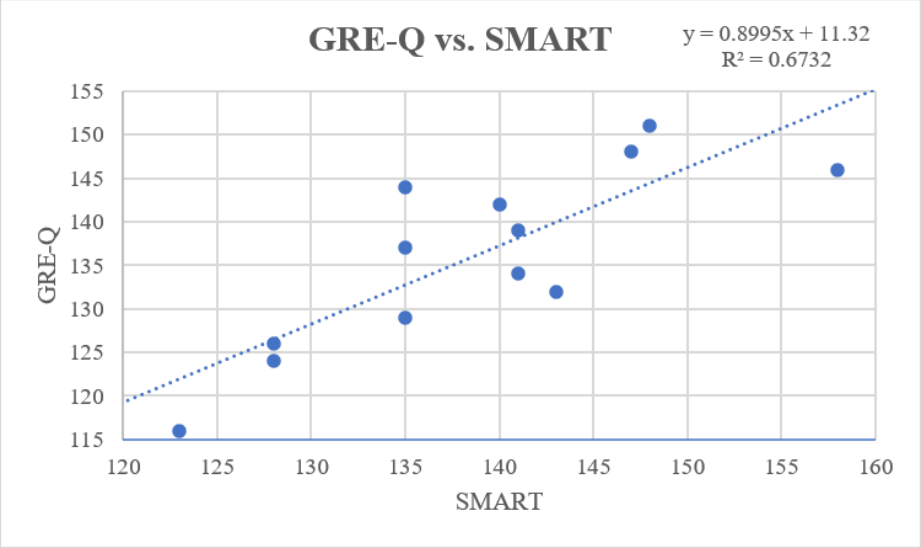
n = 38 SAT-M scores were collected. SMART correlated with SAT-M at **r = 0.766**.

SAT-M				
Min	Max	Mean	Median	St. Dev.
118	152	138.2	139	9.3



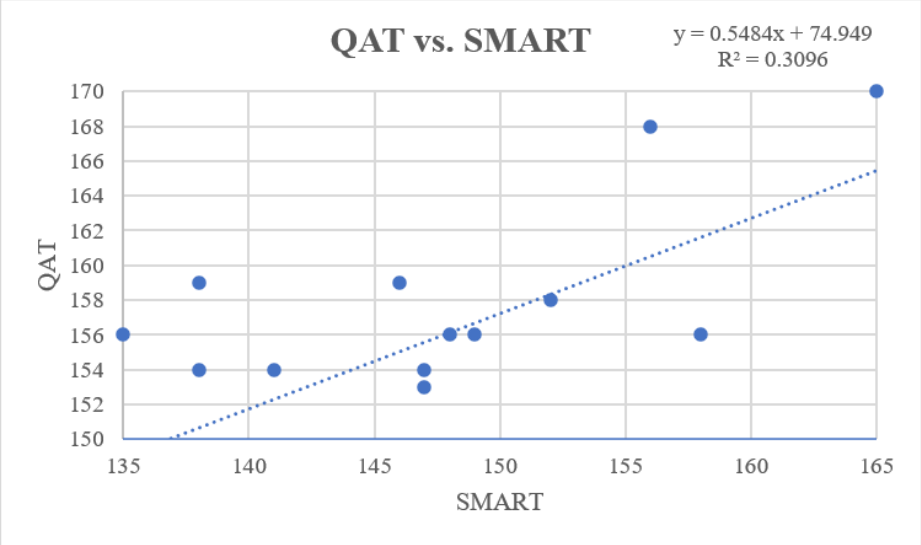
n = 13 GRE-Q scores were collected. SMART correlated with GRE-Q at **r = 0.820**.

GRE-Q				
Min	Max	Mean	Median	St. Dev.
116	151	136	137	10.0



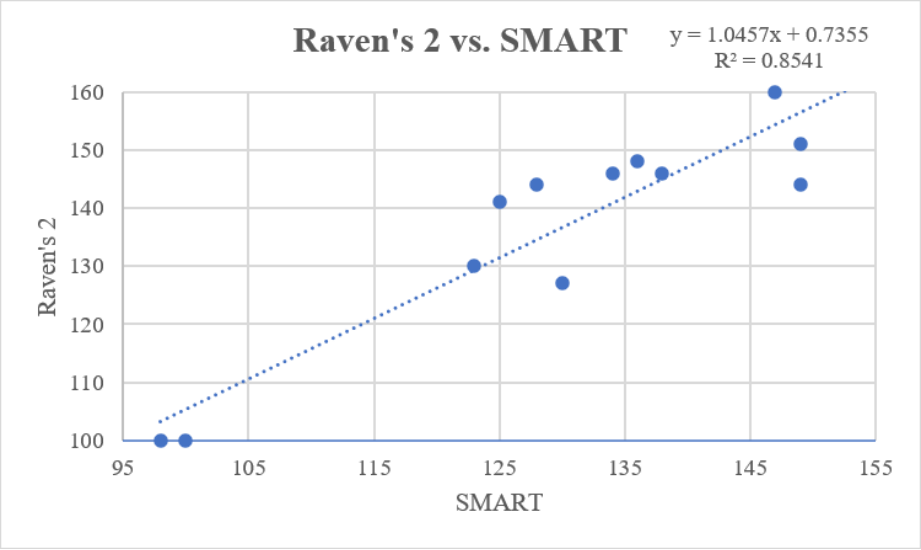
n = 15 QAT scores were collected. SMART correlated with QAT at **r = 0.556**.

QAT				
Min	Max	Mean	Median	St. Dev.
136	170	155.5	156	7.8



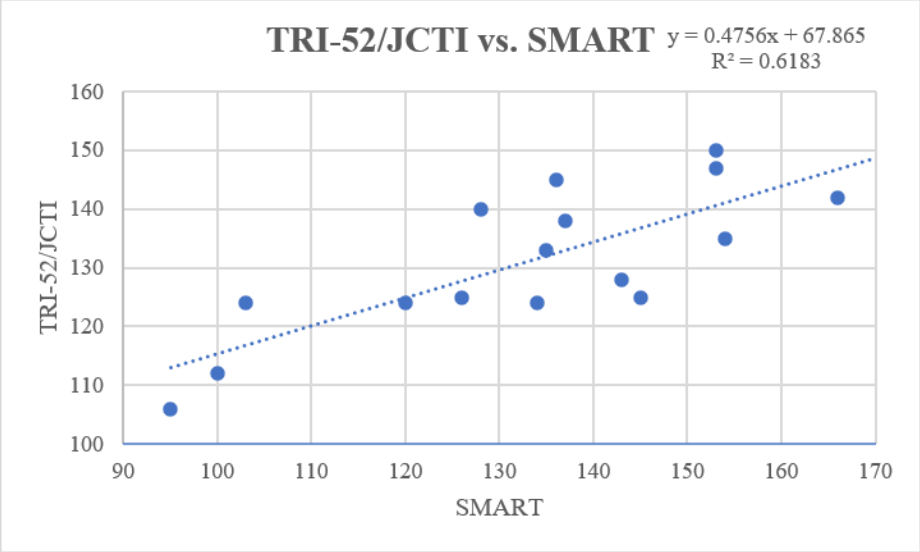
n = 12 Raven's 2 scores were collected. SMART correlated with Raven's 2 at **r = 0.924**.

Raven's 2				
Min	Max	Mean	Median	St. Dev.
100	160	136.4	144	18.3



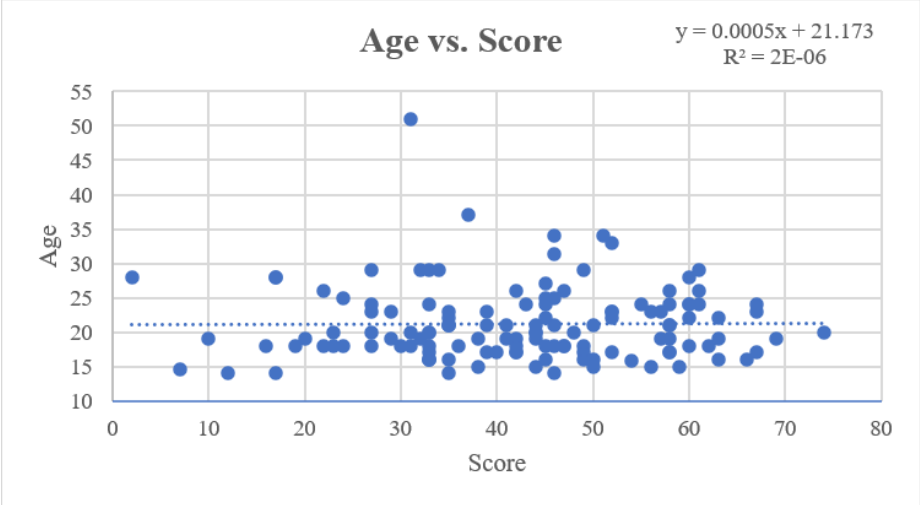
n = 21 JCTI scores were collected. SMART correlated with JCTI at **r = 0.786**.

JCTI				
Min	Max	Mean	Median	St. Dev.
106	150	132.3	135	11.5



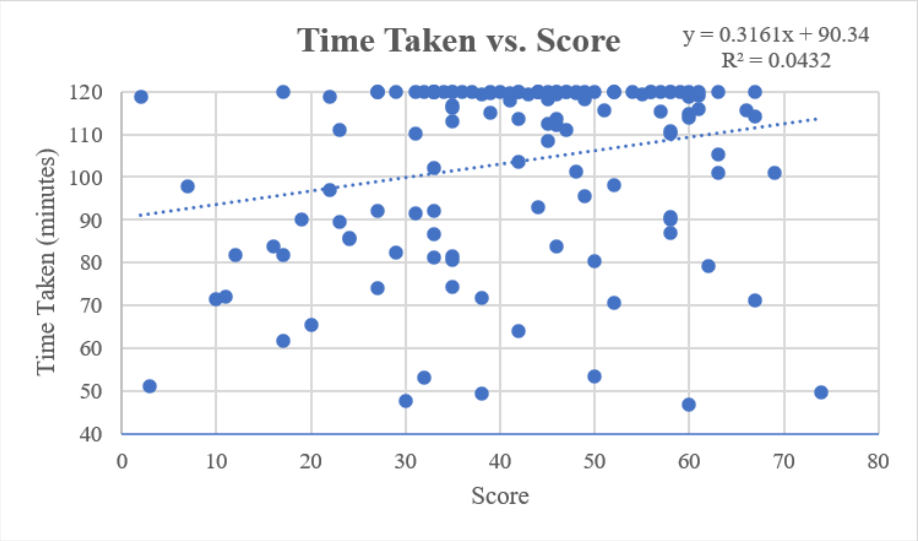
There was **no correlation** between age and score.

Age				
Min	Max	Mean	Median	St. Dev.
14	51	21.2	20	5.4



The score correlated with the time taken at **r = 0.208**.

Time				
Min	Max	Mean	Median	St. Dev.
46.8	120	104.4	115.8	20.8



Note: Tables and plots were generated only for n ≥ 10.

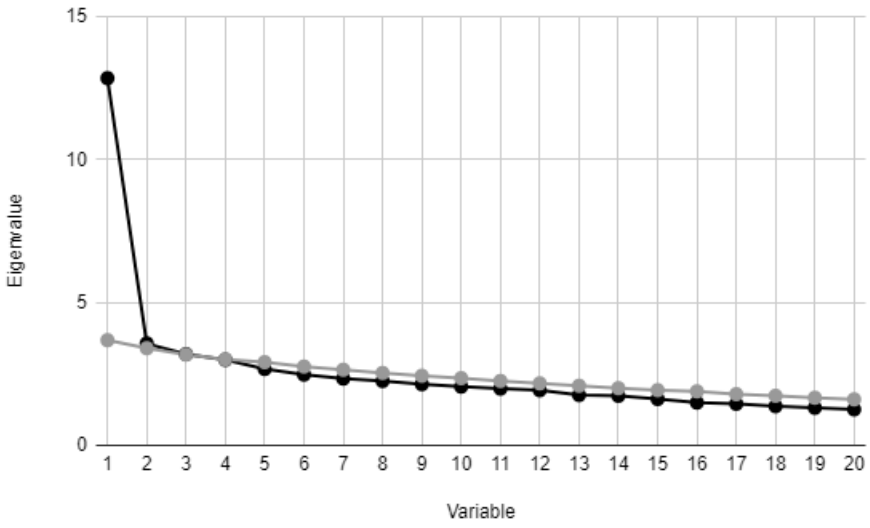
All professional quantitative (and fluid) tests used showed a moderate to high correlation with SMART, indicating that it is a valid quantitative assessment. Neither age nor the time spent on the test significantly affected performance.

V. Factor Analysis

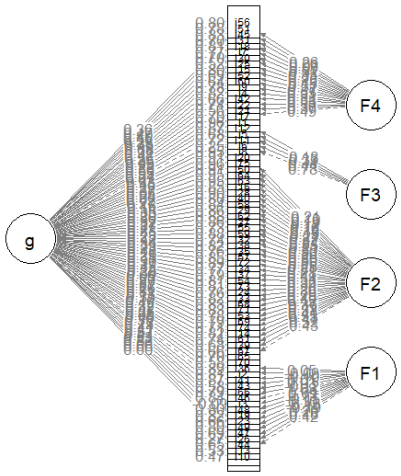
The factor structure of SMART was explored using **exploratory factor analysis** (**EFA**).

By examining the **scree plot**, we can determine the number of factors to extract. To make a more accurate decision on the number of factors, one can compare the eigenvalues to the 95th percentile estimates from parallel analysis. If an eigenvalue exceeds the parallel analysis estimate, it should be extracted. Parallel analysis generates eigenvalues from a Monte-Carlo simulated matrix created from random data. In other words, if a factor's eigenvalue exceeds the 95th percentile parallel analysis estimate, it indicates the factor's significance is beyond what could be attributed to random chance, affirming the validity of extracting that factor.

The scree plot indicates that 3 or 4 factors should be extracted using oblimin rotation. For the subsequent confirmatory factor analysis, items were assigned to the factor on which they showed the highest loading.



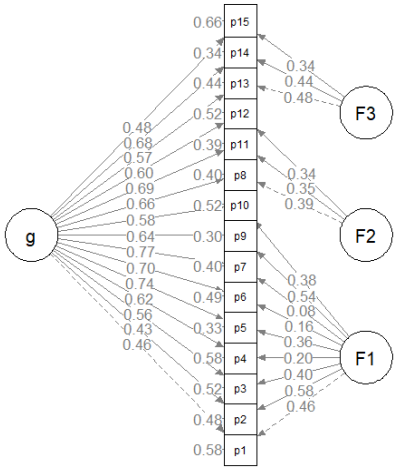
A **confirmatory factor analysis** (**CFA**) was conducted to validate this factor structure. However, the bifactor model did not fully resolve due to the excessive number of variables, making optimization and evaluation of its fit to the data nearly impossible. Nonetheless, a preliminary factor loading for SMART was calculated at 0.852.



In a second attempt at validation, the number of variables was reduced by grouping them into parcels of five items to facilitate the analysis.

The scree plot and parallel analysis showed that only one factor should be extracted with EFA – this is *g*. However, calculating a reliable *g*-loading without higher-order factors was not feasible. For convenience, three factors were extracted instead.

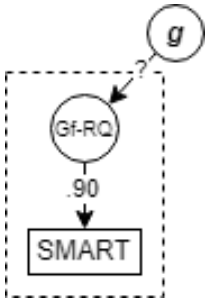
This time, the bifactor model resolved successfully. All goodness-of-fit measures were satisfactory, indicating a robust model. The factor loading was recalculated at 0.900.



Fit Measures							
Measure	cmin/df	p-value	CFI	GFI	AGFI	SRMR	RMSEA
Threshold	<3	>0.05	>0.95	>0.95	>0.80	<0.09	<0.05
Actual	1.030	0.407	0.996	0.988	0.978	0.044	0.018

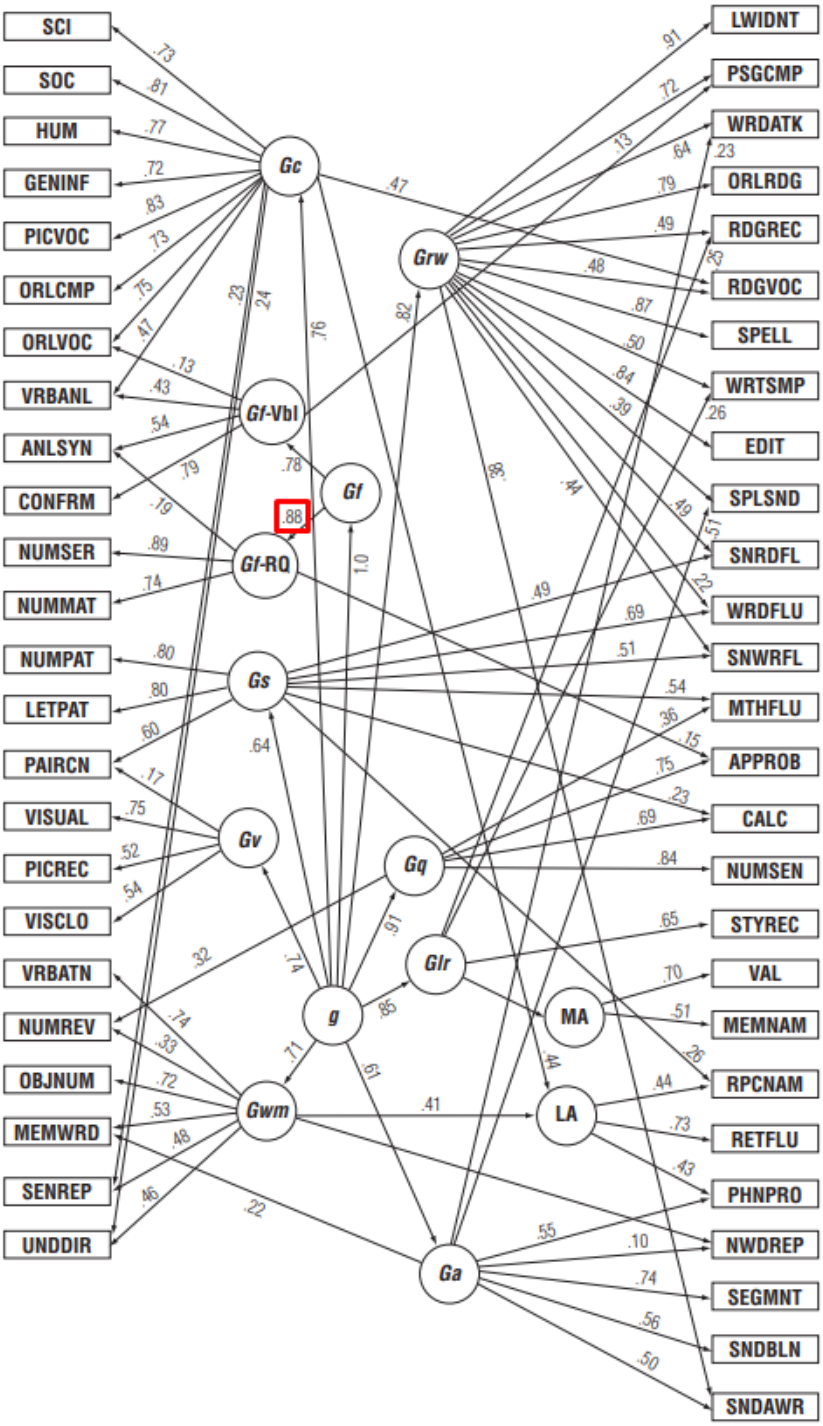
However, this loading is too high to be on *g*. Given the unidimensional nature of SMART as a quantitative test, we speculate that this factor is more likely to represent a narrower ability (Gf-RQ) than general intelligence (*g*).

If we assume that the measured loading corresponds to Gf-RQ instead of *g*, then we have the following situation:



The true loading of Gf-RQ on *g* is unknown for the current sample, but it can be estimated. Once it is estimated, the measured loading can be adjusted accordingly and rendered more representative.

In a study based on the Woodcock-Johnson IV (WJ-IV), RQ was found to load 0.88 on Gf (synonymous with *g*):

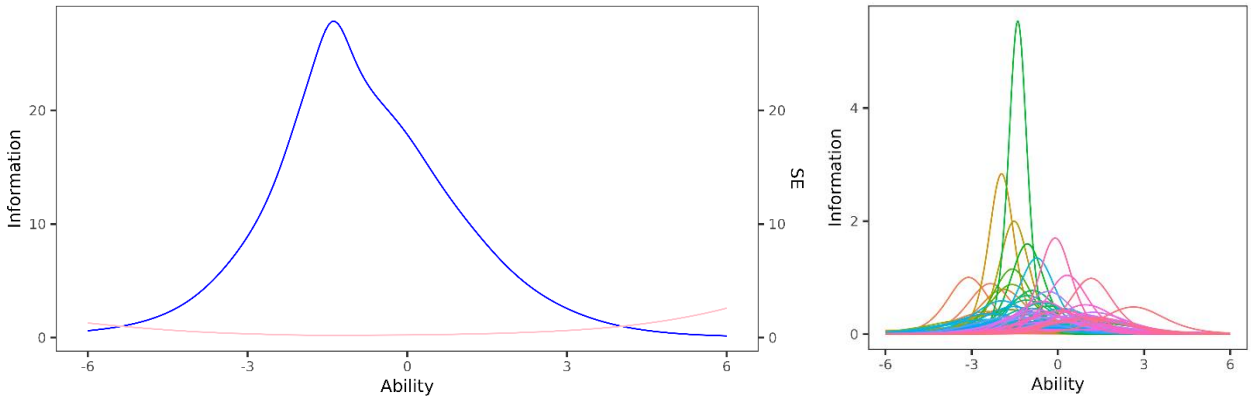


When we adjust this broad factor loading for the mean and SD of the SMART sample, it becomes 0.631 at 136.3 IQ @ 12.75 SD. Multiplying the Gf-RQ loading of SMART with the Gf-RQ correlation with g yields 0.900 × 0.631 = 0.568. This is the test's g-loading at 136.3 IQ @ 12.75 SD. After correcting for the mean and SD of the general population, we finally find that SMART has a g-loading of **0.844** at 100 IQ.

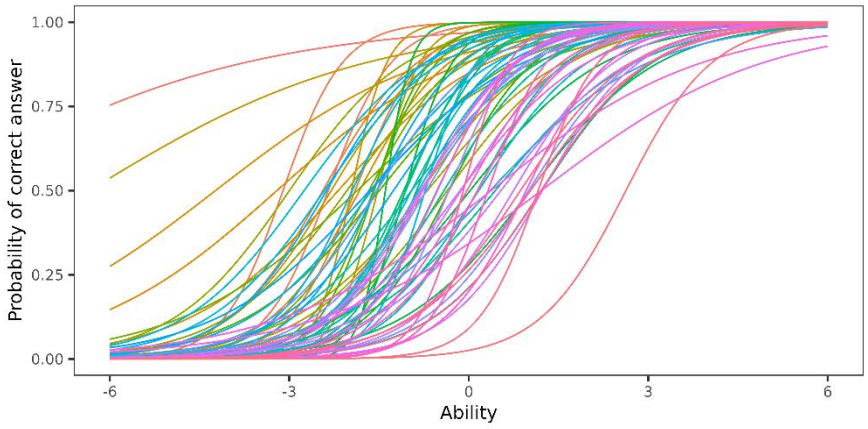
Please note that the reliability of these *g*-loading values is subject to the **sample size**. The sample size ($n = 91$ at the time when factor analysis was conducted) is **below the recommended threshold** for factor analysis, which typically requires a few hundred participants. The relatively low number of testees may **limit the accuracy** of the g-loading estimate. These findings may be confirmed with an increased sample size in the future.

VI. Item Response Theory

Item Response Theory (**IRT**) allows a nuanced analysis of items by evaluating their individual difficulty and discriminating power. The test information curve (in blue) represents the sum of the individual item information curves, revealing the test's granularity. The zone with the highest information corresponds to the zone with the lowest measurement error, making it the most accurate range of the test.



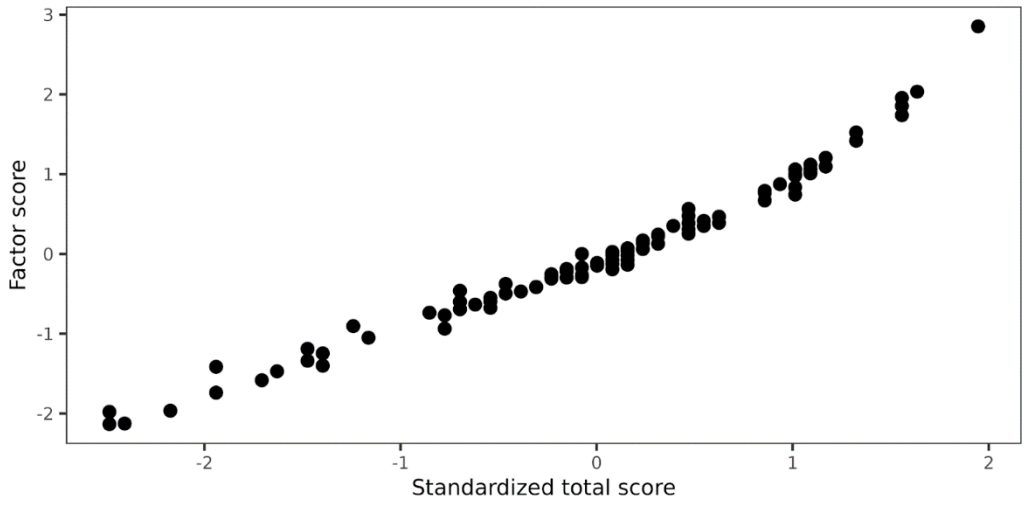
The curve's breadth across the ability spectrum, with minimal gaps, demonstrates the test's capability to discern fine differences in testee performance. Note that the floor of the test has limited discriminating power. However, this is expected as the test is not designed to measure the low ranges of ability accurately.



In IRT, the significance of an item is proportional to its discriminating ability, meaning that responses to more discriminant items have greater weight in scoring. This weighting mechanism enables a more refined scoring system, where test-takers with identical raw scores may receive different weighted scores depending on the specific items they answered correctly or incorrectly.

For example, a testee who correctly answers an easier but highly discriminant item will receive a higher factor score than a testee who correctly answers a harder but highly ambiguous item. This phenomenon commonly occurs in high-range tests (HRTs), where the more experienced testees might solve the most convoluted items but not perform as well on standard IQ tests with far lower ceilings. This is because these easy items have much higher discriminating power than the hardest HRT items, rendering the latter meaningless as they fail in their purpose of discriminating between high levels of ability.

The following figure illustrates the correlation between the weighted and raw scores, which was found to be 0.980, suggesting a strong relationship while also allowing for subtle scoring adjustments based on item discrimination. Given the complexity of calculating these factor scores on the fly, it is deemed unnecessary to substitute them for the standard raw scores.



The following table presents the 20 most difficult items identified by IRT, their discriminating abilities, and their correct rates. The hardest item presented a difficulty of **+2.63z**.

Question	Difficulty	Discrimination	**Correct Rate**
75	**2.626**	1.386	**5.49**
59	1.231	1.242	23.08
29	1.22	1.079	25.27
63	1.211	0.537	35.16
71	1.208	1.094	25.27
74	1.159	1.986	18.68
72	1.03	0.995	29.67
28	1.022	0.925	30.77
53	1.002	1.126	28.57
65	0.958	1.441	26.37
73	0.863	1.067	31.87
58	0.747	1.189	32.97
57	0.536	0.916	39.56
33	0.443	0.855	41.76
64	0.373	0.565	45.05
50	0.357	0.754	43.96
67	0.323	**2.041**	39.56
39	0.166	1.095	46.15
68	0.147	1.292	46.15
37	0.143	1.345	46.15
27	0.022	0.893	49.45

VII. Conclusion

In conclusion, the SAT Math: Advanced Rendition Test (SMART) has proven itself to be a **highly reliable** and **valid** tool for assessing advanced quantitative reasoning skills. Through meticulous construction and alignment with the original SAT-M standards, SMART has successfully extended the test ceiling, providing a more challenging assessment for high-ability testees. The comprehensive reliability measures, strong correlations with established tests, and extensive factor analysis all underscore SMART's efficacy.

Despite these strengths, it is acknowledged that the current sample size falls short of the ideal for factor analysis, which may influence the precision of the *g*-loading estimates. Future studies with larger sample sizes are anticipated to validate these findings further.

Special thanks to kronen for the advanced statistical analyses and interpretations.

Please direct any questions or comments to *u/soapyarm*.