human esources®
Your potential. Our passion.

AchieveWorks® Aptitudes

Statistical Analysis

Psychometric Report

Introduction

AchieveWorks® Aptitudes is a multi-domain cognitive assessment battery designed to help high school students gain meaningful insight into their intellectual strengths, explore career pathways that align with those strengths, and reflect on areas for continued growth. The battery comprises 11 subtests spanning eight cognitive domains — Computation, Math Reasoning, Logic, Vocabulary, Fluid Reasoning (FR), Spatial Ability, Memory, and Processing Speed (PS) — totaling 243 items administered to 2,611 students across grades 9–12 in Alabama.

This report summarizes the psychometric evidence supporting *AchieveWorks Aptitudes*, including the reliability and validity of each subtest. Overall, the battery demonstrates a strong psychometric foundation. Eight of eleven subtests meet or exceed stringent reliability standards, with several reaching excellent levels appropriate for individual-level score interpretation. Subtests that fell below reliability targets in this initial validation have already undergone item-level revision informed directly by these findings, and improved performance is anticipated as the evidence base continues to grow. The psychometric results presented here reflect a first large-scale administration under real-world conditions. The amount of support and structure during this implementation varied, and this report should therefore be considered a conservative estimate of the battery's true measurement quality compared to more structured and practiced implementations.

About the Assessment

AchieveWorks Aptitudes is designed to serve high school students as both a self-discovery tool and a career planning resource. Rather than functioning as a high-stakes diagnostic instrument, the battery is intended to help students identify their cognitive strengths, reflect on their own abilities, and explore career pathways that align with their aptitude profile. Results are reported at three developmental levels, with tailored guidance provided at each level to support continued growth regardless of where a student lands.

The battery covers eight cognitive domains through 11 subtests totaling 243 items. Five domains are measured through a single subtest each: Computation, Mathematical Reasoning, Vocabulary, Logic, and Spatial Ability. Memory, Fluid Reasoning, and Processing Speed are each assessed through two subtests, one image-based and one text-based. This dual-modality design serves two important purposes. First, it produces a more robust measure of each construct by capturing it through different formats. Second, it improves accessibility: students with reading or language differences can rely more heavily on image-based subtests, while students with visual impairments can engage fully with text-based subtests via screen reader.

All subtests use a timed, multiple-choice format with dichotomous scoring, meaning each item is scored as correct or incorrect. The assessment is designed for flexible implementation, allowing students to engage with subtests voluntarily and for time limits to vary according to need. This soft-implementation model reflects the practical realities of school settings and respects student autonomy. The section Intended Use and Appropriate Interpretation discusses how a guided and encouraging implementation by educators is expected to improve both the quality of psychometric performance and the precision of each student's career match profile.

Sample Description

The validation sample for *AchieveWorks Aptitudes* comprised 2,611 grade 9–12 students from the state of Alabama. Prior to analysis, 202 records were excluded as likely quality assurance or administrator test sessions rather than genuine student responses, ensuring that all reported metrics reflect authentic student performance.

Table 1 shows the distribution of the sample by grade and gender. Grade 9 students represent the largest portion of the sample at 64%, with female and male students roughly balanced at 33% and 31% respectively. Upper grade representation is smaller, ranging from 4% to 8% per cell. This grade distribution reflects the early adoption pattern typical of newly launched assessments and is expected to broaden as implementation expands across grade levels.

Table 1. Sample Distribution by Grade and Gender

	Grade 9	Grade 10	Grade 11	Grade 12
Female	33%	4%	5%	8%
Male	31%	5%	7%	7%

Because the battery uses a soft-implementation model, not all students attempted every subtest. Subtest-level sample sizes therefore vary, ranging from 857 students for the Spatial subtest to 1,728 for Computation. Reported psychometric metrics for each subtest are based on the students who attempted that subtest rather than the full sample. See Table A1 in the Appendix for more descriptive statistics on the subtests.

Demographic information was captured separately from the item response data to protect student privacy. So, while the demographic breakdown is shown for the entire sample, it is not currently available at a more granular level, such as scoring or other psychometric measures.

It is also worth noting that the administration conditions for this validation sample were largely informal. Anecdotal feedback from participating schools suggests that students were given varying levels of context, structure and encouragement. This is relevant context for interpreting the psychometric results: engagement levels and completion rates under more structured educator-guided implementation would be expected to yield stronger reliability estimates and more complete student profiles.

Reliability

Reliability refers to the consistency of an assessment — the degree to which it produces stable, dependable results. A reliable subtest measures its target construct in a consistent way across items, meaning that students who perform well on some items in a subtest tend to perform well on others. When reliability is high, educators and students can have greater confidence that a score reflects a genuine underlying ability rather than random variation or measurement noise.

Three reliability estimates are reported for each subtest. Cronbach's alpha (α) is the most widely recognized reliability statistic and reflects the degree to which items within a subtest are measuring the same construct. McDonald's omega (ω) is a more modern and often more accurate estimate that is preferred when items contribute unequally to the overall construct, which is common in cognitive assessments. The Spearman-Brown split-half coefficient provides an additional check by comparing performance on two halves of each subtest. Together, these three estimates provide a well-rounded picture of each subtest's consistency.

For reference, reliability values of .80 and above are considered excellent, .70 to .79 are acceptable, and values between .60 and .69 are marginal for individual-level interpretation. It is important to note that shorter subtests are inherently more difficult to make reliable, as reliability is partly a function of item count. Several *AchieveWorks Aptitudes* subtests are intentionally brief to respect student time and maintain engagement, which places a natural ceiling on achievable reliability.

Table 2. Subtest Reliability Estimates

Subtest	Items	N	Cronbach's α	McDonald's ω	Split-Half	Result
Computation	40	1,728	.93	.93	.96	Excellent
Math Reasoning	26	948	.82	.84	.82	Excellent
Vocabulary	26	1,145	.78	.80	.82	Acceptable
Logic	15	1,221	.70	.72	.69	Acceptable
FR Images	13	1,399	.65	.66	.59	Marginal
FR Text	23	1,271	.61	.65	.64	Marginal
Spatial	15	857	.70	.77	.71	Acceptable
Memory Images	14	1,446	.62	.67	.61	Marginal
Memory Text	20	1,275	.76	.79	.77	Acceptable
PS Images	16	1,151	.84	.89	.87	Excellent
PS Text	35	973	.90	.93	.94	Excellent

Several findings are worth highlighting. Computation and PS Text are the battery's strongest subtests, both achieving excellent reliability across all three estimates. PS Images and Math Reasoning also perform strongly, with Math Reasoning reliability rising to excellent levels (α = .87, ω = .89) when analysis is focused on students who engaged fully with the subtest rather than giving up partway through. This was determined by analyzing the time spent and the answering pattern over the sequence of the assessment. When the time spent was under five seconds per question and the answering pattern suddenly switched from correlational to random, the subject could be assumed to be disengaged. A comparison of results between all students and those identified as fully engaged suggests that the standard figures for Math Reasoning modestly understate true reliability among motivated students.

Eight of eleven subtests meet or exceed acceptable reliability ($\alpha \geq .70$), and the battery median McDonald's ω is .79, indicating solid overall consistency for a first large-scale administration under informal conditions.

FR Images, FR Text, and Memory Images returned marginal reliability in this validation. This is not unexpected. The FR and Memory subtests are among the shorter subtests in the battery, and reasoning constructs are inherently more difficult to measure consistently with a small number of items. All three subtests have since undergone targeted item-level revision informed directly by these findings. Improved reliability is anticipated as revised items are validated through continued administration.

Validity

While reliability tells us whether an assessment produces consistent results, validity addresses a deeper question: does the assessment actually measure what it claims to measure? A valid subtest produces scores that relate meaningfully to other measures of the same or similar constructs, and that are distinct from measures of unrelated constructs. Both of these properties were examined for *AchieveWorks Aptitudes* through an analysis of inter-subtest correlations.

Convergent validity refers to the degree to which subtests that should be related, actually are related. When two subtests measure similar or overlapping constructs, their scores should correlate positively and meaningfully. Discriminant validity refers to the opposite: subtests measuring genuinely different constructs should correlate less strongly, confirming that each subtest is capturing something distinct rather than simply re-measuring the same underlying ability.

Table 3 presents the three strongest correlates for each subtest alongside the weakest correlate, providing a snapshot of both convergent and discriminant validity across the battery.

Table 3. Convergent and Discriminant Validity: Top Correlates Per Subtest

Subtest	Correlate 1	r	Correlate 2	r	Correlate 3	r	Weakest Correlate	r
Computation	Math Reasoning	.43	FR Text	.39	Spatial	.33	Memory Images	.20
Math Reasoning	Logic	.52	Vocabulary	.46	Computation	.43	Memory Images	.18
Vocabulary	PS Text	.48	Logic	.48	Spatial	.47	Memory Images	.25
Logic	Math Reasoning	.52	Vocabulary	.48	FR Text	.47	Memory Images	.21
FR Images	PS Text	.48	Spatial	.46	PS Images	.46	Memory Images	.27
FR Text	PS Text	.48	Logic	.47	FR Images	.42	Memory Images	.23
Spatial	PS Images	.53	PS Text	.51	Vocabulary	.47	Memory Images	.20
Memory Images	PS Images	.34	Memory Text	.34	PS Text	.31	Math Reasoning	.18
Memory Text	PS Images	.49	PS Text	.48	Spatial	.41	Computation	.31
PS Images	PS Text	.63	Spatial	.53	Memory Text	.49	Computation	.27
PS Text	PS Images	.63	Spatial	.51	Vocabulary	.48	Memory Images	.31

Several validity findings are noteworthy. The strongest single correlation in the battery is between PS Images and PS Text (r = .63), providing clear evidence that the two subtests are measuring the same underlying construct through different formats. This is a textbook convergent validity result and speaks directly to the integrity of the dual-modality design.

The quantitative reasoning cluster also shows strong convergent validity. Math Reasoning and Logic correlate at r = .52, the highest correlation outside of the Processing Speed pair, and both subtests show meaningful relationships with Vocabulary and Computation. This pattern is consistent with what researchers refer to as general cognitive ability, a broad factor that tends to run through performance across many different types of cognitive tasks.

Discriminant validity is supported by the consistent finding that Memory Images is the weakest correlate for nearly every other subtest in the battery. This indicates that Memory Images is measuring something sufficiently distinct from the other constructs that it does not inflate correlations across the board. While Memory Images is the subtest most in need of psychometric refinement, its discriminant pattern is actually a meaningful finding in its own right.

Across the battery as a whole, the validity evidence supports the interpretation that *AchieveWorks Aptitudes* is measuring a meaningful and differentiated set of cognitive constructs. Subtests within the same domain relate to each other as expected, subtests across domains show moderate relationships consistent with shared general cognitive ability, and no subtest shows such high correlations with others as to suggest redundancy.

Subtest Profiles

This section provides a brief profile of each of the 11 subtests, organized by domain. Each profile summarizes what the subtest measures, its psychometric performance, and guidance on appropriate score interpretation. Reliability verdicts reference the standards introduced in the section Reliability.

Quantitative Domain

Computation

The Computation subtest measures the ability to accurately solve numerical problems across a range of arithmetic and mathematical operations. With 40 items and a sample of 1,728 students, it is the largest and most extensively validated subtest in the battery. Reliability is excellent (α = .93, ω = .93), and validity is strong, with the highest correlation to Math Reasoning (r = .43) as expected. Computation scores are appropriate for individual-level interpretation and represent one of the most dependable measures in *AchieveWorks Aptitudes*.

Math Reasoning

The Math Reasoning subtest measures the ability to apply mathematical thinking to solve problems that require reasoning and judgment rather than straightforward calculation. It is distinct from Computation in that success depends less on procedural fluency and more on conceptual understanding. With the full test group, the reliability is excellent (α = .82, ω = .84). Among students who engaged fully with the subtest (defined in the Reliability section), reliability reaches even higher (α = .87, ω = .89), making it one of the stronger measures in the battery under motivated conditions. Its strongest correlate is Logic (r = .52), reflecting the reasoning demands shared by both subtests. Math Reasoning scores are appropriate for individual-level interpretation among engaged students.

Verbal Domain

Vocabulary

The Vocabulary subtest measures the breadth and precision of a student's word knowledge, which is closely associated with verbal reasoning and general academic ability. Reliability is acceptable (α = .78, ω = .80), and the subtest shows broad validity, correlating meaningfully with Processing Speed Text (r = .48), Logic (r = .48), and Spatial (r = .47). This pattern of correlations across diverse domains reflects vocabulary's well-established relationship with general cognitive ability. Scores are suitable for individual-level interpretation with appropriate confidence intervals.

Reasoning Domain

Logic

The Logic subtest measures the ability to evaluate arguments, identify valid conclusions, and reason systematically from premises. With 15 items, it is one of the shorter subtests in the battery, which places a natural ceiling on achievable reliability. The current reliability is acceptable (α = .70, ω = .72). Validity is good, with strong correlations to Math Reasoning (r = .52), Vocabulary (r = .48), and FR Text (r = .47), confirming that Logic is capturing a meaningful reasoning signal. Items have been revised following this validation, and improved reliability is anticipated.

Fluid Reasoning Images

The FR Images subtest measures the ability to recognize and predict patterns in visual information, using image-based questions that require students to identify relationships among shapes, objects, and visual sequences. The dual-modality design makes this subtest particularly accessible for students with reading or language differences. With 13 items, reliability is currently marginal (α = .65, ω = .66), and scores are recommended for use as part of a composite profile. Validity is moderate, with meaningful correlations to PS Text (r = .48), Spatial (r = .46), and PS Images (r = .46). Items have been revised and improved reliability is anticipated in subsequent administrations.

Fluid Reasoning Text

The FR Text subtest measures the ability to recognize and predict patterns in text-based information, using sequences of words, letters, and numbers. It is designed to complement the image-based section and extends accessibility to students with visual impairments who engage via screen reader. Reliability is currently marginal (α = .61, ω = .65), and like its image-based counterpart, scores are best used as part of a broader profile rather than as standalone individual measures. Validity is good, with strong correlations to PS Text (r = .48), Logic (r = .47), and FR Images (r = .42), the last of which confirms that both fluid reasoning subtests are capturing a shared underlying construct. Items have been revised and improved reliability is anticipated.

Spatial Domain

Spatial

The Spatial subtest measures the ability to mentally visualize, manipulate, and reason about objects and their relationships in space. Reliability is acceptable (α = .70, ω = .77), and validity is moderate, with the strongest correlates being Processing Speed Images (r = .53), Processing Speed Text (r = .51), and Vocabulary (r = .47). The correlation with Processing Speed measures reflects the visual processing demands shared by both constructs. Spatial scores are usable at the individual level with appropriate

caution, and the subtest contributes meaningfully to the career matching algorithm for pathways where spatial thinking is a relevant aptitude.

Memory Domain

Memory Images

The Memory Images subtest measures the ability to encode and recall visual information. It returned the weakest psychometric performance in this validation, with marginal reliability (α = .62, ω = .67) and validity (maximum r = .34 with Processing Speed Images and Memory Text). Scores are best used as part of a broader profile rather than as standalone individual measures. This subtest has undergone the most substantial revision following this validation, with item redesign focused on improving difficulty range, discrimination, and construct coherence.

Memory Text

The Memory Text subtest measures the ability to encode and recall text-based information, including words, sequences, and verbal material. Reliability is acceptable (α = .76, ω = .79), and validity is moderate to good, with strong correlations to Processing Speed Images (r = .49), Processing Speed Text (r = .48), and Spatial (r = .41). Notably, Memory Text and Memory Images correlate at only r = .34, suggesting that visual and verbal memory are being measured as meaningfully distinct constructs rather than interchangeable forms of the same ability. This finding is consistent with well-established cognitive science research. Baddeley's widely cited Working Memory Model, for example, proposes that the memory system includes separate components for handling verbal and visual information, operating largely independently of one another. Paivio's Dual Coding Theory similarly holds that verbal and visual information are encoded and stored through distinct cognitive systems. The relatively low correlation between the two Memory subtests in *AchieveWorks Aptitudes* is therefore not a design inconsistency but a theoretically expected and meaningful result, and one that supports the value of measuring both forms of memory within the battery. Memory Text scores are suitable for individual-level interpretation within a broader profile. See the section Further Reading for more information on the Working Memory Model and Dual Coding Theory.

Processing Speed Domain

Processing Speed Images

The PS Images subtest measures the rate and accuracy with which students process and respond to visual information under timed conditions. Reliability is excellent (α = .84, ω = .89), and validity is strong, with a particularly notable correlation with PS Text (r = .63) confirming that both subtests are measuring the same underlying speed of processing construct through different formats. PS Images scores are appropriate for individual-level interpretation and contribute meaningfully to career matching for roles requiring rapid visual processing.

Processing Speed Text

The PS Text subtest measures the rate and accuracy with which students process and respond to text-based information under timed conditions. It is among the strongest subtests in the battery, with excellent reliability (α = .90, ω = .93) and broad validity across multiple domains. Its correlation of r = .63 with Processing Speed Images represents the strongest convergent validity result in the battery. PS Text scores are fully appropriate for individual-level interpretation and represent one of the most dependable and well-validated measures in *AchieveWorks Aptitudes*.

Intended Use and Appropriate Interpretation

AchieveWorks Aptitudes is designed for use in high school settings as a career exploration and self-reflection tool. It is not intended for high-stakes diagnostic or clinical applications, and results should not be used as the sole basis for consequential decisions about individual students. Used appropriately, the battery provides students and educators with a meaningful, evidence-based reference point for conversations about strengths, growth areas, and career pathways.

Score Reporting and Developmental Levels

Rather than reporting raw scores or percentiles, *AchieveWorks Aptitudes* places students into one of three developmental levels for each subtest: early, middle, and advanced. These levels correspond approximately to the lower, middle, and upper thirds of the score distribution within the norming sample, and are updated as the respondent base grows. This approach is intentionally designed to be constructive and growth-oriented. Students at every level receive tailored guidance and development recommendations, and no level carries a stigmatizing label. The three-level reporting structure is also well matched to the battery's reliability profile: while some subtests are not precise enough to support fine-grained individual score distinctions, they are sufficiently reliable to support meaningful placement into broad performance bands.

Career Matching

The career matching algorithm at the heart of *AchieveWorks Aptitudes* draws on a student's full aptitude profile across all completed subtests. The algorithm considers both the absolute level of each aptitude and the relative pattern of strengths and areas for growth across the profile, identifying career pathways where a student's cognitive strengths are most closely aligned with the demands of those careers.

An important practical consideration for educators is that the precision of career matching depends on the completeness of a student's profile. When a student completes most or all subtests, the algorithm has a rich set of variables to work with and can identify best-fit pathways with greater confidence. When only a few subtests are completed, the matching is based on less information and the results are necessarily less tailored. Educators are encouraged to support students in completing as much of the battery as possible, not because partial profiles are invalid, but because more complete profiles yield more personalized and useful career guidance.

Implementation Guidance

The psychometric results presented in this report were obtained under conditions in which the degree of instruction and encouragement varied. Research in educational assessment consistently shows that student motivation and engagement meaningfully affect performance on voluntary assessments. It is reasonable to expect that more structured, educator-guided implementation will improve both the quality of individual results and the overall psychometric performance of the battery.

Educators are encouraged to introduce *AchieveWorks Aptitudes* with context about its purpose and relevance to students' own futures, to allow adequate time for completion, and to frame participation as a valuable self-discovery opportunity rather than a performance measure. Students who understand why they are taking an assessment and who feel supported in doing so are more likely to engage thoughtfully, producing results that more accurately reflect their true abilities and that generate more meaningful career match recommendations.

Limitations and Development Roadmap

No assessment in its first large-scale validation is without limitations, and *AchieveWorks Aptitudes* is no exception. The developers are committed to transparency about current constraints and to ongoing improvement informed by psychometric evidence. The limitations described below are presented as honest context to help educators interpret results appropriately. This section is also meant to reflect the rigorous self-evaluation underlying the assessment's continued development.

Sample Characteristics

The current validation sample is drawn exclusively from grade 9 through 12 students in Alabama, with a substantial concentration of grade 9 students representing 64% of the sample. While this provides a meaningful foundation for initial validation, it means that psychometric results are most directly applicable to grade 9 students in similar school contexts. As *AchieveWorks Aptitudes* is implemented more broadly, the evidence base will naturally grow to reflect a wider range of students, strengthening the generalizability of the battery's norms and psychometric properties.

Administration Conditions

As noted in the Sample Description and Intended Use and Appropriate Interpretation sections, the validation data were collected under informal conditions with variable levels of educator guidance. This is likely to have introduced some degree of noise into the psychometric results, particularly through lower engagement and incomplete subtest participation among some students. The reliability and validity estimates reported here are therefore considered conservative, and improved performance would be anticipated under more structured implementation.

Subtest Reliability

Three subtests, FR Images, FR Text, and Memory Images, returned reliability estimates slightly below the acceptable threshold in this validation. Each of these subtests has since undergone targeted item-level revision informed directly by the findings of this analysis. The revisions addressed specific psychometric weaknesses identified through item-level and factor-analytic examination, including low item discrimination, construct drift, and item difficulty imbalance. Improved reliability is anticipated as revised items are validated through continued administration, and results for these subtests in the current version should be viewed in the context of the other measures rather than as measures meant for individual interpretation.

Construct Coverage

The dual-modality design for Fluid Reasoning, Memory, and Processing Speed reflects a deliberate effort to measure each construct through multiple formats, improving both robustness and accessibility. The finding that visual and verbal memory subtests correlate at only r = .34 confirms that these formats are capturing meaningfully distinct aspects of each construct rather than simply duplicating the same measure. As the battery matures, further examination of the relationships among subtest pairs will inform decisions about construct definitions, scoring, and profile interpretation.

Looking Ahead

AchieveWorks Aptitudes is an actively developing product. Each administration cycle generates new data that informs item refinement, norming updates, and validity research. The limitations described in this section represent the starting point of that process, not its end point. Educators and administrators who implement the battery contribute directly to its improvement, and the psychometric evidence base

will continue to strengthen as the respondent base grows and diversifies through broader implementation.

Technical Appendix

This appendix provides complete technical data for *AchieveWorks Aptitudes* for readers who wish to examine the psychometric evidence in greater detail. All metrics reported here are referenced in the body of the report and are provided in full tabular form for transparency and professional review.

A. Descriptive Statistics

All 11 subtests and their score ranges are shown. The raw scores reflect the penalty adjusted score of correct responses (+1) plus incorrect responses (-.25). The mean percent correct is simply the number of correct responses out of the total number of questions. The mean percent correct scores hover around the 50% midpoint (38.3 - 70.4) reflecting the intended target difficulty. These ranges will be used to produce a norm-referenced score in future versions of the assessment. The norm-referenced tables will be updated regularly.

Table A1. Descriptive Statistics of the Subtests

Subtest	K (items)	N (attempts)	Min Possible	Max Possible	Raw score Mean	Raw Score Standard Deviation	Mean % correct
Computation	40	1,728	-10.00	40	15.90	10.36	48.8%
Math Reasoning	26	948	-6.50	26	7.02	5.73	38.3%
Vocabulary	26	1,145	-6.50	26	7.06	5.24	39.2%
Logic	15	1,221	-3.75	15	4.24	2.95	40.8%
FR Images	13	1,399	-3.25	13	4.23	2.75	44.8%
FR Text	23	1,271	-5.75	23	5.80	3.70	38.9%
Spatial	15	857	-3.75	15	4.83	3.43	43.8%
Memory Images	14	1,446	-3.50	14	6.78	3.25	65.8%
Memory Text	20	1,275	-5.00	20	8.68	5.22	56.9%
PS Images	16	1,151	-4.00	16	10.08	4.79	70.4%
PS Text	35	973	-8.75	35	15.23	9.08	53.2%

B. Give-Up-Cleaned Reliability Estimates

For three subtests, Computation, Math Reasoning, and Processing Speed Text, reliability was also computed on a give-up-cleaned sample. This analysis censored responses submitted after a student appeared to have disengaged from the subtest, isolating the performance of students who engaged fully

with each item. The parameters for that determination were time spent on the test and the answering pattern over the sequence of the assessment. When the time spent was under five seconds per question and the answering pattern suddenly switched from correlational to random, the subject could be assumed to be disengaged. Both parameters had to be present for this assumption. Table B1 presents standard and cleaned estimates side by side.

The measures provided in the Reliability section are based on all attempts. The give-up-cleaned measures here are only for reference and to support the statement that future implementations that provide greater structure and encouragement are likely to produce better results.

Table B1. Give-Up-Cleaned Reliability: Computation, Math Reasoning, and Processing Speed Text

Subtest	Version	N	Mean % Correct	Standard Deviation	Cronbach's α	McDonald's ω	Standard Error of Measurement
Computation	Standard	1,728	49%	8.95	.93	.93	2.34
Computation	Cleaned	1,120	55%	9.32	.94	.94	2.28
Math Reasoning	Standard	948	38%	4.99	.82	.85	2.09
Math Reasoning	Cleaned	417	46%	5.93	.87	.89	2.11
PS Text	Standard	973	53%	7.23	.90	.90	2.29
PS Text	Cleaned	774	59%	6.47	.89	.90	2.16

For Computation, the give-up cleaning produces a modest reliability gain, with alpha rising from .93 to .94. For Math Reasoning the gain is more substantial, with alpha rising from .82 to .87 and omega from .85 to .89, meaningfully strengthening reliability. For Processing Speed Text, reliability is effectively unchanged by the cleaning, as the standard figures were already at excellent levels. The Standard Error of Measurement improvement for Processing Speed Text from 2.29 to 2.16 is nonetheless a meaningful gain in score precision for individual students.

C. Inter-Subtest Correlation Matrix

Table C1 presents pairwise Pearson correlations among all 11 subtests, computed on penalty-adjusted total scores. Only students who attempted both subtests in each pair are included in the respective correlation. Correlations are presented as an indicator of convergent and discriminant validity across the battery.

Table C1. Inter-Subtest Correlation Matrix

	Computation	Math Reasoning	Vocabulary	Logic	FR Images	FR Text	Spatial	Mem Images	Mem Text	PS Images	PS Text
Computation	1.00										
Math Reasoning	.43	1.00									
Vocabulary	.32	.46	1.00								
Logic	.32	.52	.48	1.00							
FR Images	.28	.37	.38	.40	1.00						
FR Text	.39	.43	.40	.47	.42	1.00					
Spatial	.33	.38	.47	.40	.46	.42	1.00				
Memory Images	.20	.18	.25	.21	.27	.23	.27	1.00			
Memory Text	.31	.32	.38	.35	.37	.40	.41	.34	1.00		
PS Images	.27	.31	.42	.37	.46	.43	.53	.34	.49	1.00	
PS Text	.30	.36	.48	.47	.48	.48	.51	.31	.48	.63	1.00

D. Methodology Notes

Reliability estimates were computed as follows. Cronbach's alpha was computed using standard item covariance methods and assumes tau-equivalence among items. McDonald's omega was derived from the first principal component of each subtest's item correlation matrix and is preferred when item loadings vary, as is typical in cognitive assessments. The percentage of variance explained by the first principal component ranges from 12% for FR Text to 34% for Processing Speed Images, reflecting meaningful variation in the degree of unidimensionality across subtests. Spearman-Brown split-half reliability was computed using an odd-even item split with the Spearman-Brown correction applied. All inter-subtest correlations are pairwise Pearson correlations on penalty-adjusted total scores, with each pair computed only on students who attempted both subtests. The give-up cleaning procedure identified and censored post-disengagement responses based on response time and pattern criteria applied at the item level.

Further Reading

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press.

Paivio, A. (1971). *Imagery and verbal processes*. Holt, Rinehart and Winston.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.