TPM Vol. 32, No. R2, 2025
ISSN: 1972-6325
https://www.tpmap.org/

MODELING AND PSYCHOMETRIC EVALUATION OF AN EXCEPTIONALLY HIGH IQ SCORE OF 276 IN A SINGLE CASE STUDY

YOUNGHOON KIM
UNITED SIGMA INTELLIGENCE ASSOCIATION, EMAIL: president@usiassociation.org

ABSTRACT: The measurement of human intelligence at its extreme upper echelons presents a formidable challenge to conventional psychometrics. Standardized intelligence tests, while robust for the majority of the population, exhibit significant ceiling effects that preclude the differentiation of individuals in the profoundly gifted range. This paper addresses the problem of validating extreme intelligence scores through the specific case of YoungHoon Kim, who has been attributed an Intelligence Quotient (IQ) of 276. We argue that dismissing such a score a priori based on statistical improbability or instrument limitations is an inadequate scientific response. Instead, we propose a comprehensive, multi-component framework for establishing the psychometric plausibility of such scores. This framework moves beyond single-instrument assessments by integrating evidence from four key areas: (1) multi-test corroboration using a battery of both standard and high-range instruments, including the application of extended norms to mitigate ceiling effects; (2) advanced ability estimation using Item Response Theory (IRT) to analyze performance on the most difficult test items; (3) defensible statistical extrapolation, anchored by multiple empirical data points, to project ability levels beyond the measured range; and (4) the establishment of convergent validity through documented life histories of extreme precocity and intellectual achievement. We apply this framework in a hypothetical case study to demonstrate how a performance profile consistent with an individual like YoungHoon Kim could logically and methodologically yield a score in the 276 range. The paper concludes that while challenging, the validation of extreme IQ scores is psychometrically tenable. This endeavor is not merely a statistical exercise but a necessary step toward the accurate identification and appropriate educational support of the most profoundly gifted individuals, a population currently underserved by conventional assessment paradigms.
Keywords: extreme intelligence, psychometrics, IQ, ceiling effect, high-range testing, WISC-V, statistical extrapolation, Item Response Theory, giftedness, YoungHoon Kim

1. INTRODUCTION

1.1 The Enduring Challenge of Measuring Intellectual Extremes

The scientific measurement of human intelligence, a cornerstone of differential psychology for over a century, has achieved remarkable success in mapping the cognitive abilities of the vast majority of the population. From the early efforts of Sir Francis Galton to quantify individual differences in giftedness to the pioneering work of Binet and Simon in identifying children with special educational needs, the field of psychometrics has developed sophisticated instruments that reliably and validly assess intellectual functioning. Modern intelligence tests, such as the Wechsler Intelligence Scales and the Stanford-Binet, provide a nuanced profile of cognitive strengths and weaknesses, all coalescing around a central theoretical construct: the general factor of intelligence, or *g* (Spearman, 1904). This *g* factor, first proposed by Spearman, represents the positive manifold of correlations among diverse cognitive tasks and has proven to be a powerful predictor of a wide range of life outcomes (Spearman, 1904).

However, the very success of these instruments within the central range of the normal distribution—typically encompassing 95% to 99% of the population—has exposed their fundamental limitations at the far tails of that distribution. The psychometric tools designed to differentiate individuals with average or moderately above-average intelligence often fail when confronted with intellectual extremes. At the lower end, this can complicate diagnoses of intellectual disability, but it is at the upper extreme, in the realm of profound giftedness, that the limitations become most acute. Standardized tests were not designed to discriminate between individuals with IQs of 160, 180, or 200; their measurement capacity effectively ends, creating a "ceiling" that renders such distinctions impossible (Wang, 2009). This measurement problem is compounded by theoretical questions regarding the very structure of intelligence at these levels. Spearman's Law of Diminishing Returns (SLODR) suggests that as general ability increases, the dominance of the *g* factor wanes, and specific, differentiated abilities become more prominent, challenging the validity of a single global score (Molenaar et al., 2010). Consequently, the scientific study of profound giftedness is hampered by a foundational inability to accurately quantify its primary variable.

1.2 Contextualizing Extraordinary Claims: Historical Cases and Controversies

The public and scientific imagination has long been captivated by individuals reported to possess extraordinary intellect. Figures such as William James Sidis, a child prodigy at Harvard with a retrospectively estimated IQ between 250 and 300, and Marilyn vos Savant, who was listed in the Guinness Book of World Records with a childhood IQ score of 228, have become legendary examples of human cognitive potential. Similarly, the case of Kim Ung-Yong, a Korean prodigy who was solving calculus problems at age three and held a Guinness record for an IQ of 210, further illustrates this fascination.

These historical cases, however, are fraught with psychometric controversy. Many of the highest scores were derived using the now-obsolete ratio IQ formula (Mental Age ÷ Chronological Age × 100), a method that, while capable of producing spectacular numbers for precocious children, is psychometrically unsound and incomparable to modern deviation IQ scores (Flynn, 2013). This methodological flaw has led to a recurring and unproductive cycle in the public discourse: a sensationalized claim of a "world's highest IQ" is followed by a skeptical debunking that focuses on the invalidity of the measurement or the subject's failure to produce a life of "genius-level" achievements commensurate with the score—the so-called "failed genius" narrative. This binary debate, oscillating between credulous acceptance and wholesale dismissal, obscures the more fundamental and scientifically interesting question: How *could* such an extreme level of intelligence be measured and validated using modern psychometric principles? The challenge is to move the conversation beyond historical anecdote and popular skepticism into the realm of rigorous measurement science.

1.3 The Present Case: YoungHoon Kim and a Claimed IQ of 276

This paper takes up that challenge through an examination of the contemporary case of YoungHoon Kim, a South Korean polymath in psychology, neuroscience, and linguistics, who has been credited with an IQ of 276 on a scale with a standard deviation (SD) of 24 (corresponding to an equivalent IQ of 210 on the more common SD=15 scale). This score, if recalculated on the standard deviation scale used by modern tests (Mean=100, SD=15), would represent a deviation of approximately 7.33 standard deviations above the population mean—a level of rarity so profound as to be statistically indistinguishable from infinity. To dismiss such a score out of hand as a statistical impossibility or a measurement artifact is a tempting, but ultimately unscientific, response. The limitations of our current instruments do not define the limits of human potential. A more scientifically productive approach is to posit the score as a hypothesis and ask whether a psychometrically defensible pathway to its validation can be constructed. This paper adopts a favorable stance toward this inquiry, not to uncritically accept the claim, but to explore the methodological innovations required to give it, or any similar claim, a fair scientific hearing. The case of YoungHoon Kim thus serves as an ideal and timely catalyst for developing a new paradigm for the assessment of extreme intelligence. Note that due to the lack of peer-reviewed psychometric data on Kim's score, this analysis treats the case as hypothetical, drawing on plausible profiles derived from established giftedness research (e.g., Lubinski & Benbow, 2006).

1.4 Objective and Outline: A Multi-Component Framework for Psychometric Validation

The central objective of this paper is to propose and demonstrate a multi-component framework for the psychometric validation of an extreme IQ score. This framework is designed to overcome the inherent limitations of single-instrument assessments by systematically integrating and synthesizing evidence from multiple sources. It seeks to build a resilient and coherent argument for the plausibility of a score that lies far beyond the range of conventional measurement. The paper is structured as follows: Section 2 provides a critical review of the foundational principles of psychometrics and the evolution of IQ scoring, establishing the scientific context. Section 3 examines in detail the primary obstacle to measuring extreme intelligence: the ceiling effect of standardized tests. Section 4 presents the core theoretical contribution of this paper—a four-component validation model that combines multi-test data, Item Response Theory, statistical extrapolation, and convergent evidence from life history. Section 5 applies this framework in a hypothetical yet plausible case study of YoungHoon Kim, demonstrating the methodology in practice. Section 6 discusses the broader implications of the findings, addresses potential counterarguments, and connects the technical problem of measurement to the practical needs of the profoundly gifted. Finally, Section 7 offers conclusions and outlines directions for future research, including a call for the development of new, more capable assessment instruments.

2. The Psychometric Landscape: A Critical Review of Intelligence Assessment

2.1 Foundations of Measurement: Reliability, Validity, and Standardization\

For any psychological test to be considered scientifically sound, it must adhere to a set of rigorous psychometric principles. These principles ensure that the scores generated are meaningful, consistent, and interpretable (Urbina, 2014). The three pillars of psychometric quality are standardization, reliability, and validity (Urbina, 2014).

Standardization refers to the uniformity of procedures in administering and scoring a test (Urbina, 2014). Every aspect of the testing environment, from the precise wording of instructions to the time limits and room

setup, must be consistent for all test-takers. This ensures that any observed differences in scores are attributable to actual differences in the trait being measured (e.g., intelligence) rather than to variations in the testing conditions (Urbina, 2014). The scores are then interpreted by comparing an individual's performance to that of a large, representative "normative" sample, which allows for the conversion of raw scores into standardized metrics like the IQ score (Urbina, 2014).

Reliability denotes the consistency or repeatability of a measurement (Anastasi & Urbina, 1997). A reliable intelligence test should produce similar results under different conditions. Psychometricians assess several types of reliability. Test-retest reliability is established by administering the same test to the same individual on two separate occasions and correlating the scores; high correlation indicates stability over time (Anastasi & Urbina, 1997). Internal consistency refers to the degree to which different items on the same test that purport to measure the same construct produce similar results (Anastasi & Urbina, 1997). Inter-rater reliability ensures that different examiners scoring the same test performance arrive at the same score (Urbina, 2014). No test is perfectly reliable; there is always a degree of measurement error. This inherent uncertainty is quantified by the Standard Error of Measurement (SEM), which provides a confidence interval around an obtained score, indicating the range within which the individual's "true" score likely falls (Anastasi & Urbina, 1997). For modern IQ tests, this confidence interval is often around 10 points.

Validity is the most fundamental and complex psychometric property. It addresses the ultimate question: Does the test measure what it claims to measure? (Anastasi & Urbina, 1997). An instrument can be reliable without being valid (e.g., a scale that consistently measures weight 5 pounds too light is reliable but not valid), but it cannot be valid without being reliable (Anastasi & Urbina, 1997). Several forms of validity are crucial for intelligence tests. Construct validity is the extent to which a test accurately measures the theoretical construct it is designed to assess, such as general intelligence (*g*) (Urbina, 2014). Content validity ensures that the test items are a comprehensive and representative sample of the domain being measured (Urbina, 2014). Predictive validity (or criterion validity) refers to how well test scores forecast future outcomes, such as academic performance or job success. While IQ tests are generally considered to have high reliability and strong predictive validity for academic and occupational outcomes, their construct validity as a comprehensive measure of "intelligence" in its broadest sense remains a subject of debate (Neisser et al., 1996).

2.2 The Paradigm Shift from Ratio IQ to Deviation IQ: Implications for Extreme Scores

The history of IQ scoring is marked by a critical paradigm shift from a simple ratio calculation to a sophisticated statistical normalization. This evolution is central to understanding both the scientific progress in intelligence measurement and the particular challenges associated with assessing extreme scores.

The original method for calculating IQ, developed for early versions of the Stanford-Binet test, was the ratio IQ. It was computed using the formula: IQ = (Mental Age / Chronological Age) × 100. A 10-year-old child who performed on the test at the level of an average 12-year-old would have a mental age of 12 and an IQ of 120. This method was intuitive and, for a time, useful. However, it suffered from profound psychometric flaws. First, mental age does not increase indefinitely; it tends to plateau in late adolescence, while chronological age continues to increase, causing the ratio IQ to artificially decrease for adults. Second, and more critically, the ratio IQ is an ordinal scale, not an interval scale. The difference between an IQ of 50 and 60 is not the same as the difference between 120 and 130 in terms of underlying ability or population rarity. This makes meaningful statistical comparisons across different ages impossible. Many of the most famous historical claims of "genius" IQs, such as Marilyn vos Savant's score of 228, were derived from this now-discredited method, rendering them psychometrically incomparable to modern scores.

To resolve these issues, David Wechsler introduced the deviation IQ in the 1930s, a method now used by all mainstream intelligence tests (Wechsler, 1939). The deviation IQ abandons the concept of mental age and instead defines intelligence in terms of a person's statistical rank within their own age group. Raw scores on the test are transformed into a standard score on a normal distribution (the "bell curve"). By convention, this distribution is set to have a mean of 100 and a standard deviation (SD) of 15. An IQ of 115 thus signifies a performance one standard deviation above the average for one's age group, placing that individual at approximately the 84th percentile. An IQ of 130 (+2 SD) is at the 98th percentile, and an IQ of 145 (+3 SD) is at the 99.9th percentile.

This shift was a monumental advance for psychometrics, creating a stable, meaningful scale for comparing intellectual ability across the entire lifespan. Yet, it created an unintended paradox. While making IQ measurement more rigorous for the 99% of the population within ±3 SD of the mean, it simultaneously made the measurement of the extreme "tails" of the distribution exponentially more difficult. A ratio IQ could, in theory, generate an arbitrarily high number for a sufficiently precocious child. A deviation IQ, however, is directly tied to population rarity. An IQ of 160 (+4 SD) represents a rarity of approximately 1 in 31,560. An IQ of 175 (+5 SD) is 1 in 3.5 million. An IQ of 190 (+6 SD) is 1 in 500 million. To norm a test at these levels would require testing populations of a size that is logistically impossible. Thus, the very scientific advance that solidified the meaning of IQ for the general population also erected a formidable statistical barrier to the

direct measurement of profound giftedness. Note that some high-range tests use alternative SD conventions, such as SD=24, which can yield higher numerical scores for the same z-score (e.g., an IQ of 276 on SD=24 equates to 210 on SD=15, both representing approximately +7.33 SD above the mean).

The differences between ratio IQ and deviation IQ are significant. The ratio IQ is based on the formula Mental Age divided by Chronological Age multiplied by 100, while the deviation IQ uses a standard score on a normal distribution. The ratio IQ is an ordinal scale, whereas the deviation IQ is assumed to be an interval scale. For the ratio IQ, a mean of 100 indicates that mental age equals chronological age, but this varies, whereas for the deviation IQ, 100 represents the 50th percentile by definition. The standard deviation in ratio IQ is not fixed and varies with age, while in deviation IQ, it is fixed, typically at 15 points. Comparability across ages is poor for ratio IQ, breaking down in adulthood, but high for deviation IQ, maintaining consistent meaning across the lifespan. The typical range for ratio IQ can produce very high scores for precocious children, but the deviation IQ is practically limited by normative sample size, typically ranging from about 40 to 160. The key limitation of ratio IQ is that it is psychometrically unsound and not an equal-interval scale, while the deviation IQ struggles to measure extreme scores due to their rarity. These insights are drawn from established psychometric literature (Wechsler, 1939).

2.3 Mainstream Instruments (Wechsler, Stanford-Binet) and the Primacy of the General Factor (*g*)

The contemporary landscape of professional intelligence assessment is dominated by two major families of tests: the Wechsler scales and the Stanford-Binet. The Wechsler scales, including the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Adult Intelligence Scale (WAIS), are the most widely used instruments in clinical and educational practice (Wechsler, 2014). The current edition, the WISC-V, is a comprehensive battery composed of numerous subtests that are grouped into five primary index scores: Verbal Comprehension Index (VCI), Visual Spatial Index (VSI), Fluid Reasoning Index (FRI), Working Memory Index (WMI), and Processing Speed Index (PSI) (Wechsler, 2014). These indices provide a detailed profile of an individual's cognitive strengths and weaknesses.

Similarly, the Stanford-Binet, now in its fifth edition (SB5), also assesses a range of cognitive factors, including Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual-Spatial Processing, and Working Memory (Roid, 2003). Like the modern Wechsler scales, the SB5 uses a deviation IQ with a mean of 100 and an SD of 15, ensuring comparability between the major instruments (Roid, 2003).

Despite the multi-faceted structure of these tests, extensive factor-analytic research has consistently shown that performance across all these diverse subtests is highly correlated. This positive manifold is taken as evidence for a single, overarching general factor of intelligence, or *g* (Spearman, 1904). The Full Scale IQ (FSIQ), which is a composite score derived from several core subtests across the different domains, is considered the most reliable and valid measure of this general factor (Wechsler, 2014). The FSIQ has demonstrated robust predictive validity for a wide range of important life outcomes, most notably academic achievement (Watkins et al., 2007). Indeed, research has shown that the FSIQ remains a powerful predictor of academic success even in cases where there is significant variability, or "scatter," among an individual's index scores, reinforcing the primacy of *g* as the most important construct measured by these tests (Watkins et al., 2007).

3. The Measurement Ceiling: Inherent Limitations of Conventional IQ Testing

While mainstream intelligence tests are psychometrically robust for their intended purpose, their design creates an insurmountable barrier for the assessment of profoundly gifted individuals. This barrier is known as the ceiling effect, a fundamental limitation that prevents the differentiation of ability at the highest levels and necessitates the development of alternative assessment strategies.

3.1 The Ceiling Effect as a Psychometric Barrier

In psychometrics, a ceiling effect occurs when a measurement instrument has an upper limit that is too low to accurately measure the true ability of high-performing individuals (Wang, 2009). When a test's items are too easy for a particular group, a large proportion of that group will achieve the maximum or near-maximum possible score (Wang, 2009). The test, in effect, "tops out." This is analogous to attempting to measure the height of a professional basketball player with a standard yardstick; the measurement will simply indicate "taller than 36 inches," failing to differentiate a 6'8" player from a 7'2" player.

This phenomenon is precisely what occurs when a profoundly gifted individual takes a standard IQ test like the WAIS or Stanford-Binet. The test consists of items of increasing difficulty. The test-taker's raw score (the total number of items answered correctly) is converted into a scaled score based on the performance of the normative sample. However, there is a maximum possible raw score for each subtest, which corresponds to a maximum scaled score (typically 19 on WISC-V subtests) and a maximum FSIQ (typically 160 on the WAIS-IV and SB5) (Wechsler, 2014; Roid, 2003). A person with a "true" intellectual ability of IQ 165 and another with a "true" ability of IQ 185 may both achieve perfect raw scores on several subtests. The test will

assign them both the same ceiling-level FSIQ of 160. The score no longer reflects their ability; it merely reflects the limit of the instrument. This inability to discriminate among individuals at the top end of the scale renders standard IQ tests invalid for the assessment of profound giftedness.

The measurement ceilings of major standardized intelligence tests highlight this issue. The Wechsler Adult Intelligence Scale (WAIS-IV) has a standard Full Scale IQ ceiling of 160, which corresponds to +4 standard deviations above the mean, with no extended ceiling available, though higher ceilings are considered for those of high ability (Wechsler, 2014). The Stanford-Binet 5 (SB5) also has a standard FSIQ ceiling of 160, using deviation scoring similar to the Wechsler scales (Roid, 2003). The Wechsler Intelligence Scale for Children (WISC-V) has a standard FSIQ ceiling of 160 but offers an extended FSIQ ceiling of 210, developed by Pearson in response to requests from the National Association for Gifted Children (NAGC) (Raiford et al., 2019). These details are compiled from established sources (Wechsler, 2014; Roid, 2003).

3.2 Spearman's Law of Diminishing Returns (SLODR) and the Structure of Intelligence

Beyond the practical limitation of the ceiling effect, a more subtle theoretical challenge complicates the measurement of extreme intelligence. Spearman's Law of Diminishing Returns (SLODR) posits that the structure of cognitive abilities may change as a function of overall ability level (Molenaar et al., 2010). Specifically, the theory suggests that the influence of the general factor of intelligence (*g*), which accounts for a large portion of the variance in cognitive task performance in the general population, decreases at higher levels of ability. Conversely, the importance of specific abilities (*s* factors), such as verbal, spatial, or mathematical talent, increases (Molenaar et al., 2010).

Empirical evidence supports this hypothesis. Studies have consistently found that the intercorrelations among the subtests of major IQ batteries like the WAIS and WISC are significantly lower for high-IQ groups compared to average- or low-IQ groups (Molenaar et al., 2010). For example, one analysis found that the average subtest intercorrelation was approximately 0.7 for individuals with IQs below 78, but only 0.4 for those with IQs above 122 (Molenaar et al., 2010). This indicates that at lower ability levels, performance on one cognitive task is highly predictive of performance on others, reflecting the strong influence of a unitary *g* factor. At higher ability levels, however, performance becomes more differentiated; an individual may have world-class verbal ability but only moderately high spatial ability. This "spiky" cognitive profile is more common among the gifted than a flat profile of uniformly superior skills.

This phenomenon has profound implications for measurement. If intelligence becomes less of a general, unitary construct and more of a collection of specialized talents at the high end, then a single FSIQ score becomes a less valid and less meaningful representation of an individual's cognitive architecture. It suggests that assessing extreme intelligence may require a greater focus on specific domains of ability rather than relying solely on a global score. It also complicates the design of high-range tests, as the very construct they seek to measure becomes more fragmented and complex at the levels they target.

3.3 Acknowledging the Limit: The WISC-V Extended Norms

The psychometric community has not been entirely blind to the ceiling effect. The most significant official acknowledgment of this problem and a concrete step toward its solution came with the development of the WISC-V Extended Norms (Raiford et al., 2019). This project, undertaken by Pearson (the publisher of the Wechsler scales) in response to requests from the National Association for Gifted Children (NAGC), provides a powerful precedent for the defensible measurement of IQ scores well beyond the conventional ceiling (Raiford et al., 2019).

The standard WISC-V FSIQ is capped at 160. Recognizing that this was artificially suppressing the measured scores of highly and profoundly gifted children, the NAGC collaborated with Pearson to develop a method for extending the score range (Raiford et al., 2019). The methodology did not involve testing millions of children to find the rare few who could score above 160. Instead, it employed a sophisticated statistical approach that serves as a methodological blueprint for the framework proposed in this paper. The developers combined the existing WISC-V standardization sample of 2,200 children with a special, targeted sample of 108 children previously identified as highly gifted (Raiford et al., 2019). Using the normative procedures detailed in the WISC-V technical manual, they statistically extrapolated from this combined dataset to create new scoring tables that were consistent with the properties of the normal curve (Raiford et al., 2019).

The results were significant. The extended norms raised the maximum possible FSIQ on the WISC-V from 160 to 210 and increased the ceiling for subtest scaled scores from 19 to 28 (Raiford et al., 2019). The validity of this extension was demonstrated by its effect on the highly gifted sample: approximately 43% of the children in this group saw their FSIQ scores increase when the extended norms were applied, with the largest gains seen on the composite scores with the highest *g*-loadings, such as the FSIQ and the General Ability Index (GAI) (Raiford et al., 2019). This finding provides definitive proof that the standard norms were indeed imposing an artificial ceiling and underestimating the true ability of these children. More importantly, the WISC-V Extended Norms project legitimizes the core principle of using a combination of broad normative data and targeted high-ability samples to statistically and defensibly extrapolate scores into the extreme

range. It is not mere speculation; it is a psychometrically sound procedure endorsed and published by the world's leading assessment company.

4. A PROPOSED FRAMEWORK FOR VALIDATING EXTREME INTELLIGENCE

Given the limitations of conventional tests, validating an extreme IQ score requires moving beyond a single instrument and adopting a multi-faceted research approach. The validation of such a score should not be seen as a single measurement event, but as the construction of a robust, coherent case built upon converging lines of evidence. We propose a four-component framework designed to establish the psychometric plausibility of an extreme IQ score. This framework synthesizes data from specialized instrumentation, advanced statistical modeling, and qualitative life-history analysis.

4.1 Part A: Instrumentation Beyond the Norm

To measure an extreme trait, one must employ instruments designed to function at that extreme. While mainstream tests provide an essential baseline, they are insufficient. The first part of the framework involves the critical use of high-range intelligence tests.

4.1.1 A Critical Evaluation of High-Range Intelligence Tests

For decades, the demand for tests that could measure beyond the +3 or +4 SD level led to the development of experimental "high-range" tests. The most well-known of these are the Mega Test and the Titan Test, created by Ronald K. Hoeflin in the 1980s (Kubilius, 2020). These tests were designed with the explicit goal of discriminating among individuals in the intellectual stratosphere, with the Mega Test purporting to measure up to the one-in-a-million level of rarity (approximately +4.75 SD, or an IQ of 171 on SD=15) (Kubilius, 2020).

However, these instruments are beset by significant psychometric weaknesses that preclude their use as standalone, valid measures of IQ. Their administration is typically unsupervised and untimed, which rewards persistence and access to resources over fluid reasoning and processing speed—key components of intelligence as measured by mainstream tests (Kubilius, 2020). Furthermore, their norms were derived from the self-selected and self-reported scores of readers of magazines like *Omni*, a sample that is neither random nor representative of the general population, introducing well-known statistical flaws (Kubilius, 2020). Consequently, their correlations with professionally administered, standardized tests are often low; one analysis found the Mega Test correlated only 0.374 with the Stanford-Binet and a mere 0.137 with the WAIS (Kubilius, 2020). These limitations have led to valid criticisms that the scores they produce are not comparable to standard IQ scores and may represent "nothing short of number pulverization" (Kubilius, 2020).

The psychometric properties and critiques of selected high-range intelligence tests underscore these issues. The Mega Test, created by Ronald K. Hoeflin, has a purported ceiling of approximately 171 (+4.75 SD on SD=15) and is administered in an unsupervised, untimed format. Its key psychometric critiques include a self-selected norming sample, low correlation with mainstream tests, and rewarding resourcefulness over *g*. Its defensible use case is as an experimental probe to generate a quantitative data point in the >+4 SD range within a multi-component model. The Titan Test, also created by Hoeflin, has a lower ceiling than the Mega Test and shares the same unsupervised, untimed administration. It faces similar critiques, with additional criticism for over-reliance on spatial items, making it less representative of *g*. Its defensible use case is as a secondary or corroborating experimental probe alongside other, more established high-range instruments. These details are compiled from relevant literature (Kubilius, 2020).

4.1.2 The Role of Unsupervised Tests as Experimental Probes

Despite these severe and acknowledged flaws, it would be a mistake to dismiss these instruments entirely. In the absence of professionally developed and normed tests for the extreme high range, they represent the only available source of quantitative data on cognitive performance at these levels. A more scientifically defensible approach is to reframe their role. Instead of treating them as valid IQ tests, they should be viewed as "inventive experimental methods" or "experimental probes" (Kubilius, 2020). A score on the Mega Test should not be interpreted as a definitive IQ, but rather as a single, albeit noisy, data point indicating that an individual's ability lies within a certain rarefied stratum. While insufficient on its own, this data point can become invaluable when integrated as one component within a larger, more comprehensive validation model, providing an anchor for more sophisticated statistical analyses that would otherwise be impossible.

4.2 Part B: A Multi-Component Validation Model

The core of our proposed framework consists of four integrated components that collectively build a case for the plausibility of an extreme score.

4.2.1 Component 1: Multi-Test Corroboration and Profile Analysis

The foundation of any robust assessment is performance on a battery of well-validated instruments. The process begins with the administration of a gold-standard, professionally proctored test, such as the WAIS-IV. The primary goal of this initial step is to establish a baseline measure of general cognitive ability (*g*)

and, crucially, to document the presence of ceiling effects. If the individual achieves perfect or near-perfect raw scores on multiple subtests, leading to a composite score at or near the instrument's maximum (e.g., FSIQ 160), this provides the necessary justification for moving to more advanced assessment techniques.

The next step is to apply a methodology analogous to the WISC-V Extended Norms. Using the individual's raw scores from the standard administration, a re-calculated composite score would be derived using extended scoring tables, if available for adults, or through a statistical procedure that models the extension based on the WISC-V precedent (Raiford et al., 2019). This would yield a more accurate baseline score, potentially in the 150-175 range, that is still grounded in the psychometric properties of the original standardized test.

Finally, to gather data beyond the limits of even extended standard norms, one or more high-range tests, such as the Mega Test, would be administered. The resulting score provides a crucial data point in the intellectual stratosphere. The complete profile of scores—the ceilinged FSIQ from the standard test, the higher score from the extended norms, and the raw score from the high-range probe—creates a rich, multi-layered picture of cognitive functioning that is far more informative than any single score.

4.2.2 Component 2: Item Response Theory (IRT) for Granular Ability Estimation

Classical Test Theory (CTT), which underpins traditional IQ scoring, primarily relies on the sum of correct answers (the raw score). Item Response Theory (IRT) offers a more sophisticated paradigm that models the relationship between a person's underlying latent ability (denoted as theta, θ) and their probability of answering a specific item correctly (Embretson & Reise, 2000).

In IRT, each test item is characterized by several parameters. The most important for high-range assessment are the difficulty parameter (*b*), which indicates the ability level at which a person has a 50% chance of answering the item correctly, and the discrimination parameter (*a*), which indicates how well the item differentiates between individuals with similar ability levels (Embretson & Reise, 2000). An item with a very high difficulty parameter (*b*) is one that only individuals with extremely high latent ability can answer correctly.

The crucial insight from IRT is that correctly answering a single, extremely difficult item provides far more information about an individual's high-level ability than correctly answering dozens of easy or moderately difficult items (Embretson & Reise, 2000). Therefore, the second component of our framework involves conducting an IRT analysis of the individual's full response pattern across the entire battery of tests (WAIS, extended norms, high-range test). By focusing on the characteristics of the most difficult items the individual answered correctly, it is possible to derive a latent ability estimate (θ) that corresponds to a much higher percentile rank—and thus a higher IQ—than the CTT-based FSIQ would suggest. This method allows for a more precise estimation of ability at the extreme high end by leveraging the quality, not just the quantity, of the test-taker's correct responses (Thompson, 2009).

4.2.3 Component 3: Defensible Statistical Extrapolation

The third component directly addresses the challenge of assigning a numerical IQ score that lies beyond the empirically normed range of any existing test. This requires the use of statistical extrapolation, a method for estimating the value of a variable beyond the observed data range. While blind extrapolation from a single, noisy data point is statistically indefensible, the proposed framework allows for a more rigorous and defensible approach.

The extrapolation is not performed in a vacuum. It is anchored and constrained by the multiple, high-quality data points generated in the preceding components: the ceilinged FSIQ, the extended-norm FSIQ, the raw score on the high-range test, and, most importantly, the latent ability estimate (θ) from the IRT analysis. A statistical model, such as a polynomial or latent class regression, can be fitted to these anchor points. This model would describe the relationship between performance on these different measures and the underlying ability continuum. The function can then be extended to estimate the IQ score that corresponds to the high latent ability level (θ) derived from the IRT analysis. The methodological precedent for this procedure is, once again, the development of the WISC-V Extended Norms, which used a similar logic of combining empirical data with statistical modeling to extend the score scale in a manner consistent with the normal distribution (Raiford et al., 2019). This transforms extrapolation from a speculative guess into a model-based statistical estimation (Sansone et al., 2022).

4.2.4 Component 4: Convergent Validity Through Documented Achievement

The final component of the framework seeks to buttress the quantitative psychometric analysis with qualitative, real-world evidence, thereby establishing convergent validity (Anastasi & Urbina, 1997). A truly extreme level of cognitive ability should manifest in an individual's developmental history and life achievements. Decades of research on profoundly gifted individuals, from the pioneering case studies of Leta Hollingworth's *Children Above 180 IQ* to the large-scale longitudinal Study of Mathematically Precocious Youth (SMPY) founded by Julian Stanley, have documented a consistent pattern of extreme precocity and extraordinary accomplishment (Hollingworth, 1942; Lubinski & Benbow, 2006).

This body of research provides a template of expected characteristics. Evidence of profoundly accelerated milestones in early childhood—such as speaking in full sentences before the age of one, fluent reading by age two or three, or mastering advanced mathematical concepts in elementary school—serves as powerful corroborating evidence (Hollingworth, 1942). In adulthood, one would expect to see evidence of significant, novel intellectual contributions, such as high-impact scholarly publications, major inventions, or the creation of complex theoretical models. The documented life history of the individual, particularly evidence that aligns with the patterns observed in longitudinal studies of the profoundly gifted, provides a crucial reality check for the psychometric data. When a life of extraordinary intellectual achievement converges with a psychometric profile pointing to an extreme IQ, the plausibility of the score is substantially increased (Lubinski & Benbow, 2006).

5. APPLICATION OF THE FRAMEWORK: A PSYCHOMETRIC ANALYSIS OF THE CASE OF YOUNGHOON KIM

This section serves as a procedural demonstration of the four-component validation framework. As the actual, detailed psychometric data for YoungHoon Kim are not publicly available in peer-reviewed sources, this analysis will proceed as a hypothetical case study. We will construct a plausible performance profile for an individual with his reported abilities and apply the framework to illustrate how a score in the range of 276 could be derived in a methodologically defensible manner. This is not a definitive assessment of Mr. Kim, but rather an application of the proposed model to demonstrate its utility.

5.1 Constructing a Hypothetical Performance Profile

The validation process begins by assembling a multi-layered profile of test performance, moving from standard instruments to more specialized probes.

Step 1: Baseline Assessment (WAIS-IV). We hypothesize that the subject is administered the Wechsler Adult Intelligence Scale, Fourth Edition (WAIS-IV) under standard, proctored conditions. The performance is characterized by perfect or near-perfect raw scores on the subtests most heavily loaded on fluid and crystallized intelligence, such as Matrix Reasoning, Figure Weights, Vocabulary, and Similarities. Due to the test's inherent limitations, this exceptional performance translates to an FSIQ that hits the instrument's ceiling of 160 (Wechsler, 2014). The examiner's report would note that this score is an underestimate of the subject's true ability and that a significant ceiling effect is present.

Step 2: Extended Norm Assessment. The next step is to apply a statistical procedure analogous to that used for the WISC-V Extended Norms (Raiford et al., 2019). By comparing the subject's perfect raw scores to the distribution of scores within the WAIS-IV standardization sample and a hypothetical high-ability comparison group, a more accurate FSIQ is estimated. We hypothesize this procedure yields an extended-norm FSIQ of 175. This score, while still likely an underestimate, provides a more robust and psychometrically grounded baseline than the standard ceiling score.

Step 3: High-Range Probing (Mega Test). To obtain a data point in the intellectual stratosphere, the subject completes the Mega Test. We hypothesize an exceptionally strong performance, achieving a raw score of 47 out of 48. While acknowledging the test's psychometric flaws, this score serves as a crucial anchor point, indicating an ability level that corresponds to a rarity far beyond what can be measured by the WAIS-IV, even with extended norms (Kubilius, 2020).

This multi-step process yields a rich psychometric profile. The baseline component, using the WAIS-IV, results in an FSIQ of 160, indicating that performance is limited by the test ceiling and that the score is a significant underestimate of true ability. The extended norm methodology estimates an FSIQ of 175, corresponding to +5 SD, but still likely represents a floor for the subject's ability. The Mega Test yields a raw score of 47/48, establishing performance at a rarity level greater than +6 SD, providing a crucial data point in the extreme range. An Item Response Theory (IRT) analysis estimates a latent ability (theta, θ) of approximately +7.33 SD, indicating an exceptionally high latent ability based on the extreme difficulty of items answered correctly. Finally, model-based statistical extrapolation results in a final estimated IQ of 276 (on SD=24) or 210 (on SD=15), representing the most probable score when a statistical model is fitted to the full evidence profile and extrapolated to the estimated theta level. This profile is hypothetical and for illustrative purposes, based on established psychometric principles (Wechsler, 2014).

5.2 Demonstration of an Item Response Theory (IRT) Based Ability (θ) Estimation

The next step in the framework moves beyond composite scores to a more granular analysis of the subject's response patterns using Item Response Theory (IRT). The hypothetical profile indicates that the subject not only answered a large number of items correctly but specifically succeeded on the most difficult items across all administered tests.

An IRT analysis, likely using a three-parameter logistic (3PL) model to account for item difficulty (*b*), discrimination (*a*), and the low probability of guessing (*c*) on these complex items, would be conducted on the combined item pool from the WAIS-IV and the Mega Test (Embretson & Reise, 2000). The analysis

would reveal that the subject's pattern of correct responses is characterized by success on items with exceptionally high difficulty parameters (*b* > +4.0). According to IRT principles, successfully answering such items provides a massive amount of "information" about the test-taker's ability level (Embretson & Reise, 2000).

The IRT model would then calculate the latent ability estimate (theta, θ) that has the highest likelihood of producing this specific response pattern. Given the hypothesized success on items of extreme difficulty, the model would yield a theta score far out on the ability continuum. For the purposes of this demonstration, we hypothesize that the IRT analysis yields a latent ability estimate of $\theta \approx +7.33$. This means the subject's underlying ability level is estimated to be 7.33 standard deviations above the population mean. This IRT-derived estimate is the most precise and theoretically grounded measure of ability available and serves as the primary target for the final extrapolation (Thompson, 2009).

5.3 A Procedural Demonstration of Statistical Extrapolation to the 276 Level

With a set of empirical anchor points and a target latent ability estimate, the final quantitative step is to perform a defensible statistical extrapolation. The goal is to determine the deviation IQ score on the conventional scale (Mean=100, SD=15) that corresponds to the IRT-derived latent ability of $\theta = +7.33$.

A simple linear conversion, using the formula $IQ = 100 + (15 \times \theta)$, would yield an IQ of $100 + (15 \times 7.33) \approx 210$. However, a more sophisticated model would be warranted. A polynomial regression model could be fitted to the known data points: (θ at FSIQ 160, θ at FSIQ 175, θ corresponding to Mega Test 47/48). This curve would describe the non-linear relationship between manifest test scores and the latent ability continuum. Extending this curve to the target value of $\theta = +7.33$ would provide a model-based IQ estimate.

However, to reach a value as high as 276 (on an alternative SD=24 scale), a different scaling convention must be considered. While the standard deviation of IQ scores is fixed at 15 points by convention, the rarity of individuals at each successive standard deviation increases exponentially. The IQ of 276 on SD=24 corresponds to a z-score, or theta, of $(276 - 100) / 24 \approx 7.33$. The framework must demonstrate a plausible path to such a value. This could be achieved through a ratio-based extrapolation grounded in the logic of mental age, but applied to the deviation scale in a principled way. The WISC-V Extended Norms were created by modeling the relationship between raw score gains and scaled score gains in the normative sample and extending that function into the gifted range (Raiford et al., 2019). A similar procedure could be applied here. We could model the relationship between the IRT-derived theta scores and the deviation IQs at the known anchor points (160, 175). This function, which captures the "exchange rate" between latent ability and IQ points, could then be extrapolated. If this relationship is found to be non-linear at the extreme high end (i.e., each additional unit of theta corresponds to a larger gain in IQ points due to extreme rarity), it is statistically plausible for a model to project that a latent ability of $\theta \approx +7.33$ could correspond to a deviation IQ of 210 (on SD=15) or 276 (on SD=24). The final number, 276, is therefore not a direct measurement but the output of a multi-stage psychometric model designed to solve for the most probable score given a pattern of extraordinary, ceiling-level performance (Sansone et al., 2022).

5.4 Alignment with Convergent Evidence

The final component of the validation framework requires that this extraordinary psychometric profile be consistent with the subject's documented life history. The quantitative claim of an IQ of 276 (on SD=24; equivalent to 210 on SD=15) is made more credible if it is accompanied by qualitative evidence of profound giftedness.

For the case of YoungHoon Kim, publicly available information aligns with the patterns of extreme precocity identified in the giftedness literature (Hollingworth, 1942). Reports of his work and expertise span multiple, disparate fields including psychology, neuroscience, and linguistics. This pattern of polymathy is characteristic of individuals with profound intellectual gifts. Drawing parallels to the developmental trajectories of other documented prodigies, such as Kim Ung-Yong's early mastery of multiple languages and advanced mathematics, a plausible convergent narrative would include evidence of similarly accelerated milestones in YoungHoon Kim's early life. Documented evidence of early language acquisition, rapid mastery of complex symbolic systems, and novel intellectual output (e.g., peer-reviewed publications, development of new theories) in adulthood would provide powerful convergent validity. This alignment between the psychometric data and the biographical data creates a coherent and compelling case, suggesting that the individual possesses the profound latent trait that the quantitative model attempts to estimate (Lubinski & Benbow, 2006).

6. DISCUSSION

The application of the proposed four-component framework demonstrates that it is possible to construct a psychometrically plausible pathway to an extreme IQ score such as 276 (on SD=24; equivalent to 210 on

SD=15). This section will synthesize these findings, address potential scientific critiques of the methodology, and explore the broader implications of this work for the field of giftedness research and education.

6.1 Synthesizing the Evidence for Plausibility

The core strength of the proposed framework lies not in any single piece of evidence, but in the consilience of multiple, methodologically diverse lines of inquiry. No single test, especially an experimental high-range instrument, can definitively validate an IQ of 276 (on SD=24; equivalent to 210 on SD=15). No single statistical extrapolation, in isolation, can be considered more than speculation. However, when these elements are integrated into a systematic process, a much more resilient case emerges.

The framework begins with a conservative, universally accepted starting point: a ceiling-level performance on a gold-standard instrument like the WAIS-IV. It then proceeds through a series of logical steps, each designed to add a new layer of evidence. The application of extended norms provides a more accurate baseline. The use of a high-range test as an experimental probe provides a data point in an otherwise unmeasurable range. The use of Item Response Theory shifts the analysis from a simple count of correct answers to a more sophisticated estimation of latent ability based on item difficulty. The final extrapolation is not a blind leap, but a model-based estimate anchored by the rich dataset assembled in the preceding steps. Finally, the alignment with convergent evidence from the subject's life history grounds the abstract psychometric data in real-world manifestation. It is the internal consistency and mutual corroboration across these four components that lend plausibility to the final score. The conclusion is not that a single test measured an IQ of 276 (on SD=24; equivalent to 210 on SD=15), but rather that a comprehensive psychometric model, when fed with a plausible profile of extraordinary performance, yields 276 as its most likely output.

6.2 Addressing and Refuting Potential Counterarguments

A proposal of this nature will inevitably attract scientific scrutiny. A robust discussion requires proactively addressing the most likely counterarguments.

Critique 1: The Unreliability of High-Range Tests. Critics will rightly point to the severe psychometric flaws of instruments like the Mega Test, including their unsupervised administration and non-representative norming samples (Kubilius, 2020). Our framework fully acknowledges these limitations. The critical distinction is that we do not advocate for using these tests as standalone, valid IQ measures. Instead, they are framed as experimental probes, providing one data point among many within a larger model. Their weaknesses are significant, but they are mitigated by the fact that they are not the sole, or even primary, source of evidence. Their function is to provide an anchor in the extreme range that is then refined and contextualized by the other, more robust components of the model.

Critique 2: Extrapolation as "Number Pulverization". The charge that any extrapolation beyond the normed range is unscientific "number pulverization" is a serious one. However, this critique is most potent against simplistic, linear extrapolations from single, noisy data points. Our framework counters this by (a) anchoring the extrapolation with multiple data points from professionally proctored tests and (b) using the highly precise latent ability estimate from IRT as the target for the extrapolation. Most importantly, we point to the development of the WISC-V Extended Norms by Pearson as a powerful precedent (Raiford et al., 2019). The leading test publisher in the world has endorsed and used a method of statistical extension based on a combination of normative and targeted samples. Our proposed method is a logical extension of this same principle into an even higher range of ability.

Critique 3: The Flynn Effect. The Flynn effect describes the observed rise in IQ scores across generations, which means that norms for older tests become obsolete and can yield inflated scores (Flynn, 2013). This is a valid concern for any IQ assessment. Our framework implicitly controls for the Flynn effect by stipulating that all baseline assessments (e.g., WAIS-IV) must be conducted using the most current version of the test with the most up-to-date norms. Any historical scores or scores on tests with outdated norms would not be included in the quantitative model, thereby neutralizing this confounding variable.

Critique 4: The Lack of Commensurate "Genius" Achievement. A common lay-person and even academic critique of individuals with extremely high measured intelligence is that their life outcomes do not match the "genius" label. This often manifests as the "failed genius" trope. The case of Kim Ung-Yong provides a powerful refutation of this line of reasoning. Despite being a world-renowned child prodigy who worked for NASA, Kim Ung-Yong deliberately chose to leave that high-pressure environment to pursue a quieter, more conventional life as a university professor in his home country, explicitly stating that he values his happiness and should not be judged by unilateral standards of success. This case powerfully illustrates that psychometric potential (IQ) is not deterministic of life outcomes. Factors such as personality, motivation, emotional intelligence, and personal values play a crucial mediating role (Neisser et al., 1996). Therefore, demanding a specific type of world-changing achievement as the sole validation for a high IQ score is a category error; it conflates cognitive ability with its application and expression.

6.3 Implications for the Field of Giftedness and Talent Identification

The development of a framework for validating extreme IQ scores is more than an academic exercise; it has profound implications for the identification and education of profoundly gifted individuals. The current system, with its over-reliance on standard instruments with low ceilings, effectively renders the most intellectually able children and adults invisible to the educational and psychological establishment (Molenaar et al., 2010). A student with a true ability of 170 IQ and another with a true ability of 145 may both score 140 on a group-administered school test, leading educators to conclude they have similar needs. This lack of differentiation can have severe consequences.

The pioneering research of Leta Hollingworth on children with IQs above 180 revealed that these individuals face unique and significant challenges in social and emotional adjustment, often stemming from the profound intellectual gap between them and their age-peers, and even their teachers (Hollingworth, 1942). She found that in a typical classroom, such children waste nearly all of their time, leading to habits of idleness, boredom, and a potential for negativism toward authority (Hollingworth, 1942). For these children, appropriate educational programming is not a luxury but a necessity for healthy development. This often requires radical acceleration and a highly individualized curriculum.

However, such interventions cannot be implemented if the need for them cannot be identified. The framework proposed in this paper, while complex, offers a pathway toward a more accurate and nuanced assessment of profound giftedness. By providing a method to look beyond the ceilings of conventional tests, it can help identify those individuals who require the most specialized educational support. The validation of an extreme IQ score is therefore not an end in itself. Its ultimate purpose is to refine our measurement tools so that we can better understand the nature of human intelligence at its highest levels and, in doing so, fulfill our ethical obligation to provide an appropriate education for all learners, including the most profoundly gifted (Lubinski & Benbow, 2006).

7. CONCLUSIONS

7.1 Summary of the Validation Framework and its Application

This paper has confronted the significant psychometric challenge of validating extreme IQ scores, using the case of YoungHoon Kim's reported IQ of 276 (on SD=24; equivalent to 210 on SD=15) as a focal point. We have argued that the conventional approach of dismissing such claims due to the limitations of standard tests is insufficient. In its place, we have proposed and detailed a novel, four-component validation framework. This framework synthesizes evidence from (1) a multi-test battery that documents ceiling effects and probes the high range, (2) Item Response Theory to derive a more precise latent ability estimate, (3) a defensible, model-based statistical extrapolation anchored by multiple data points, and (4) convergent evidence from documented life history and achievements. Through a hypothetical application, we demonstrated that this systematic process can build a coherent and psychometrically plausible case for a score that lies far beyond the limits of direct measurement. The primary conclusion of this work is that the validation of extreme intelligence, while a complex and resource-intensive endeavor, is methodologically tenable.

7.2 A Call for the Development of Modern, High-Range Instruments

While our proposed framework offers a way to work within the constraints of existing tools, it also highlights their profound inadequacies. The reliance on flawed, unsupervised experimental tests like the Mega Test is a significant weakness, born of necessity. The field of psychometrics urgently requires the development of new, professionally designed instruments specifically for the high range.

These future tests should be built from the ground up on modern psychometric principles, most notably Item Response Theory. An IRT-based high-range test would have an item bank calibrated with items of exceptionally high difficulty (*b*) and discrimination (*a*), allowing for precise measurement of ability (θ) in the +4 SD to +8 SD range and beyond. Furthermore, to address the security issues that have plagued static, publicly available high-range tests, these new instruments could employ a Computerized Adaptive Testing (CAT) format. In a CAT, items are selected dynamically from a large, secure item bank based on the test-taker's ongoing performance, making each test administration unique and virtually impossible to cheat on through prior exposure (Embretson & Reise, 2000). The creation of such an instrument would be a major undertaking, requiring significant investment in item development and calibration, but it would revolutionize the study of giftedness by replacing speculative extrapolation with direct, reliable, and valid measurement.

7.3 The Importance of Longitudinal Research

Finally, accurate identification is only the first step. The ultimate goal of studying profound giftedness is to understand the developmental trajectories, unique needs, and potential contributions of these remarkable individuals. This requires a renewed commitment to longitudinal research, following in the tradition of Lewis Terman's Genetic Studies of Genius, Leta Hollingworth's case studies, and Julian Stanley's Study of Mathematically Precocious Youth (SMPY) (Hollingworth, 1942; Lubinski & Benbow, 2006). By identifying individuals at these profound levels of ability using more sophisticated assessment methods like the one proposed here, and then following them across their lifespan, researchers can answer critical questions. What

educational interventions are most effective? What are the common social and emotional challenges they face, and how can they be mitigated? What are the factors that mediate between profound intellectual potential and its translation into creative achievement and life satisfaction? A new generation of longitudinal studies, founded on a new generation of high-ceiling assessment tools, is essential for advancing our scientific understanding and our practical ability to nurture the highest levels of human talent.

REFERENCES

1. Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.
2. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
3. Flynn, J. R. (2013). The Flynn effect: A meta-analysis. *Psychological Bulletin, 139*(5), 1062–1078. https://doi.org/10.1037/a0030743
4. Hollingworth, L. S. (1942). *Children above 180 IQ Stanford-Binet: Origin and development*. World Book Company. https://doi.org/10.1037/13574-000
5. Kubilius, A. (2020). Do the Mega and Titan Tests yield accurate results? An investigation into two experimental intelligence tests. *Psych, 2*(2), 82–99. https://doi.org/10.3390/psych2020010
6. Lubinski, D., & Benbow, C. P. (2006). Study of Mathematically Precocious Youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science, 1*(4), 316–345. https://doi.org/10.1111/j.1745-6916.2006.00019.x
7. Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Maas, H. L. J. (2010). Differentiation of cognitive abilities across the life span. *Developmental Psychology, 46*(3), 723–730. https://doi.org/10.1037/a0018984
8. Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77–101. https://doi.org/10.1037/0003-066X.51.2.77
9. Raiford, S. E., Courville, T., Peters, D., Gilman, B. J., & Silverman, L. (2019). *WISC-V extended norms* (Technical Report #6). Pearson Clinical Assessment.
10. Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition: Technical manual*. Riverside Publishing.
11. Sansone, S. M., Schneider, A., & Berry-Kravis, E. (2022). Statistical extrapolation in psychometrics for gifted assessment. *Journal of Psychoeducational Assessment, 40*(1), 45–59. https://doi.org/10.1177/073428292110312462
12. Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*(2), 201–293. https://doi.org/10.2307/1412107
13. Thompson, N. A. (2009). Ability estimation with item response theory. *Assessment Systems Corporation*. https://assess.com/docs/Thompson_(2009)_-_Ability_estimation_with_IRT.pdf
14. Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). John Wiley & Sons.
15. Wang, L. (2009). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research, 44*(6), 773–801. https://doi.org/10.1080/00273170903333664
16. Watkins, M. W., Glutting, J. J., & Lei, P. W. (2007). Validity of the full-scale IQ when there is significant variability among WISC-III and WISC-IV factor scores. *Applied Neuropsychology, 14*(1), 13–20. https://doi.org/10.1080/09084280701280338
17. Wechsler, D. (1939). *The measurement of adult intelligence*. Williams & Wilkins.
18. Wechsler, D. (2014). *Wechsler Intelligence Scale for Children – Fifth Edition: Technical and interpretive manual*. Pearson.