

THE IN A HURRY SERIES

Intelligence

in a Hurry

What it is, how it is measured, and the fights about it

Max Brocklesby

BOOKSINAHURRY.COM

Contents

The Whole Thing in One Page 4

Why You Should Care 6

The Core Ideas 8

How It Actually Works 22

What People Get Wrong 32

Use It 38

Terms 43

Go Deeper 46

Notes and Sources 48

Bibliography 55

The Whole Thing in One Page

An intelligence test cannot see intelligence. It sees a person choose an answer, assemble blocks, remember a sequence, define a word or discover a rule. The score comes later, after tasks, instructions, time limits, scoring rules and a comparison group have turned those performances into a number.

The false public image is a tank inside the head. Everyone possesses an amount; the test inserts a dipstick; the result reveals the level. That picture drives both worship and rejection of IQ. The better model is an inference. Psychologists use *intelligence* for recurring differences in how people learn, reason, solve unfamiliar problems and use knowledge across tasks. No test samples all of that. A good test samples enough of it, under controlled conditions, to support a useful estimate with stated uncertainty.

The estimate is possible because cognitive performances are positively correlated. People who do well on one demanding task tend, on average, to do well on others. Factor analysis compresses this positive manifold into a general dimension, usually called *g*, while broad abilities such as acquired knowledge, fluid reasoning, processing speed, working memory and visual-spatial skill preserve further structure. The positive manifold is one of the most replicated patterns in differential psychology. Its cause is less settled. A factor can organise variation without being one organ, one gene or one mental substance.

An IQ score is therefore engineered rather than discovered whole. Raw performances are converted through norms, often to a scale centred on 100 with a standard deviation of 15. Reliability asks how consistently the score behaves. Validity asks what interpretation and use the evidence supports. Fairness asks whether access, task content, prediction, errors and consequences work acceptably for the people affected. A test can answer one question well and be unfit for another. Every score also carries measurement error, even when the report prints a confident integer.

Scores survive because they predict. Broad cognitive performance is

associated with learning, educational attainment and many forms of work performance. It can help identify patterns of need. Yet prediction changes probabilities across groups; it does not write one person's future. Knowledge, motivation, personality, health, opportunity, discrimination and chance remain in the outcome. A distribution is not a biography.

Brains and genes contribute, but neither supplies a shortcut around behaviour. Intelligent performance depends on distributed, developing neural systems. Genetic influences are highly polygenic and operate through bodies and environments. Heritability describes variation in a studied population under its conditions, not a personal percentage or a claim of permanence. Schooling, health, toxins and social circumstances can change measured performance. Population averages have risen, levelled and fallen across periods too short for genetic evolution to explain.

The fights begin when one link is made to do another's work. A measured group difference is treated as its own causal explanation. A reliable rank is treated as permission to exclude. Eugenic institutions converted disputed classifications into hereditary identities. Critics then sometimes answer those abuses by denying every useful measurement. Both shortcuts hide the chain.

The final link is easy to miss. A score that changes teaching, diagnosis, employment, expectation or stigma changes the conditions under which later ability will appear. Measurement samples a life, then may help shape the next sample.

That is the book.

Why You Should Care

In work published in 1905, Alfred Binet and Théodore Simon placed brief problems before children. Follow an instruction. Name an object. Repeat digits. Compare two weights. The immediate question was not who possessed the superior mind. Schools needed a more disciplined way to identify children who might require different teaching. The evidence consisted of performances under observation, followed by a judgement about support.

That modest procedure became an administrative technology. Descendants of those tasks have helped diagnose intellectual disability, identify giftedness, classify soldiers, select employees and sort pupils into different educational routes. They have also supplied eugenics, racial hierarchy and exclusion with the appearance of measurement. A score can open a specialist classroom or close a door. It can reveal a broad difficulty, then be mistaken for the person who displayed it.

You should care because intelligence testing appears wherever uncertainty about human capability meets scarce opportunity. A school cannot observe every future lesson before deciding who needs help. An employer cannot watch an applicant solve ten years of problems. A clinician cannot infer the source of learning difficulty from conversation alone. A well-designed assessment provides a disciplined sample. Used properly, it can replace status, confidence and vague impression with evidence. Used badly, it gives an unsupported judgement a decimal point.

Even an unchanged score can hide an extraordinary amount of learning. Imagine a child who scores 100 at ten and again at fifteen. The second result does not say that five years of education achieved nothing. The comparison group has grown older too. The child may know much more while holding the same relative position. This is one reason IQ attracts bad arguments: a number that looks like an amount often describes a standing. Interpreting it also requires its error range, the suitability of the test and the circumstances in which it was taken. The integer begins the interpretation. It

does not finish it.

Then comes the question most people are trying to reach: does intelligence matter? Yes, though neither camp gets the simple answer it wants. Broad cognitive scores predict learning and are associated with grades, training and performance in many jobs. The relationship is useful enough that abolishing tests does not abolish selection. It may return selection to interviews, references, accent, family advocacy and resemblance to the selector. Yet predictive validity is not destiny. High scorers fail, lower scorers excel, and any life contains knowledge, persistence, social skill, health, support, discrimination and luck.

The genetic argument matters because it is a test of whether the reader can keep levels separate. Cognitive differences are substantially heritable in many studied populations. Schooling, nutrition, toxins, disease and social conditions also alter cognitive performance. These statements do not cancel each other. Heritability describes observed variation under conditions. It does not award ownership of a trait to nature or reveal what would happen under different conditions.

Group comparisons show the cost of losing that discipline. A difference between sample means is treated as proof of an inherited cause. Heritability within groups is imported into a comparison between groups. Social race labels are used as though they were clean genetic populations. Or every finding is dismissed because racists once used tests. Each move replaces a scientific question with allegiance.

This book offers a better route. Keep person, conditions, performance, score, model, prediction and decision distinct. Ask what evidence supports each transition. Then inspect the return path: who receives more teaching, a reduced curriculum, a diagnosis, an opportunity or a stigma because the number was believed?

The result is larger than a view on IQ. It is a method for judging every institution that turns human variation into a score and every claim that the score has already decided what someone may become.

The Core Ideas

A Pattern, Not a Substance

Give a large group several different tasks: learning an unfamiliar rule, holding steps in mind, finding a pattern or explaining a difficult paragraph. The rankings will vary. Yet, across such studies, people who perform well on one task tend to perform well on others. Some learn faster, preserve more of a problem while working and use knowledge more effectively across different demands. Intelligence research begins with that recurring pattern.

The noun makes the pattern sound like an object. Psychologists cannot remove intelligence from a skull, weigh it and put it back. They infer a latent trait from behaviour: a proposed source of consistency across tasks and occasions. This is not unique to intelligence, but the inference is unusually exposed to context because performances depend on what the person has learned, what the task demands and what condition the person is in.

A serviceable definition combines learning, reasoning, problem solving and adaptation to new demands. None of these occurs in a vacuum. Novel reasoning uses learned symbols and strategies. Vocabulary reflects years of exposure as well as the capacity to learn from it. A person who acquires knowledge quickly may later look more capable partly because that knowledge makes further learning easier. Intelligence is therefore neither stored information nor a content-free processor. It is expressed through their interaction.

This separates ability from achievement without pretending the border is sealed. An achievement test asks how much mathematics, reading or biology has been learned. An ability test tries to sample capacities that support learning across domains. Yet matrices still reward familiarity with abstract diagrams, and vocabulary remains one of the stronger indicators of broad cognitive performance. Test labels describe emphasis, not purity.

Capacity is also different from the performance available today. Sleep loss, pain, anxiety, medication, sensory impairment, language mismatch, fatigue,

misunderstanding or refusal can reduce what appears in the room. There is no direct route to the untouched capacity behind those influences.

Assessors can improve access, standardise conditions, note barriers and compare the result with other evidence. They cannot observe a person outside every state and history.

Nor is intelligence the whole of a good mind. Curiosity affects what receives attention. Conscientiousness affects what gets finished. Creativity can generate possibilities that a tightly scored item never invites. Rationality also depends on goals, reflection and concern for truth rather than processing capacity alone. Social judgement, courage, practical knowledge and moral character matter in situations where a timed puzzle contributes little. These qualities may correlate with intelligence without becoming synonyms for it.

The task determines what counts as success. Speeded items reward rapid output as well as accuracy. Spoken answers require hearing and language production. Construction tasks add visual and motor demands. Multiple choice limits expression and permits guessing. Such features are not dirt around a pure measure. They are the path through which the construct becomes observable.

What can be estimated is a recurring pattern in a person's performances. It is not a quantity of personhood.

The Correlations That Hold the Field Together

Charles Spearman began with a stubborn observation. Pupils who did well in one school subject tended to do well in others, and the same broad tendency appeared across tests that looked quite different. Vocabulary, spatial reasoning, number series, memory span and processing tasks do not measure one identical act. Their results are nevertheless usually positively correlated.

This positive manifold is the field's load-bearing finding. Without it, one overall score would be no more defensible than averaging shoe size with musical pitch. Because the tasks share variation, a composite can summarise something that travels across them. The correlations are

incomplete, which is why one score never exhausts the profile.

Factor analysis turns the correlation table into a smaller structure. A general factor captures variance spread widely across tasks. Broad factors capture additional clustering among, for example, acquired knowledge, reasoning, visual processing, memory and speed. Narrow factors and task-specific demands remain below them. The method is a disciplined compression of observations.

Compression is not a causal explanation. A railway map can reveal that many lines meet at one station without explaining why the city grew there. In the same way, g can organise covariance without showing that one mental reservoir feeds every task. The factor has no anatomical address merely because it sits at the top of a diagram.

Several causal families remain live. Common-cause accounts propose broad capacities that influence many performances. Process-overlap theories emphasise executive and attentional demands recruited across otherwise different tasks. Developmental mutualism begins with several partly independent differences, then lets them reinforce one another: easier early learning produces more knowledge and better strategies, which improve later learning. Network accounts treat general performance as an emergent property of coordinated systems rather than one central engine.

These theories do not always make sharply different predictions in ordinary datasets. The positive manifold is therefore better established than any single story about why it exists. Saying that g has been proved confuses a statistical pattern with its mechanism. Saying that g has been debunked because no single mechanism has won confuses an unsettled explanation with a missing pattern.

The battery also helps create the structure it can detect. A test dominated by verbal tasks will reveal verbal distinctions in detail. Add spatial, speed, memory and quantitative tasks and the hierarchy changes. Factor analysis cannot recover an ability the battery never sampled, and researchers can obtain different emphases by changing tasks, age ranges, models and

scoring choices. Replication concerns the broad structure, not one eternal list of factors.

Individual profiles supply the boundary. A person may combine strong verbal knowledge with slower output, strong spatial reasoning with ordinary memory span, or exceptional expertise with unremarkable general scores. Injury, education, development and ageing can alter parts of the system differently. Shared variance is substantial rather than total.

How a Test Builds a Score

A score begins long before anyone sits down. Test developers define the construct and purpose, choose tasks, pilot items, remove weak or unfair ones, decide how responses will be credited, and recruit a norming sample. Administration is standardised so that differences in instructions, timing and prompts do not become invisible differences in the measure. The polished booklet hides years of decisions.

In a block-design task, the examinee arranges coloured blocks to match a printed pattern, sometimes against the clock. Accuracy and speed can both enter the raw result. Imagine earning twelve points: the number alone says little until it is compared with results for the relevant age band. A norm table converts it to a scaled score, which can then contribute to a composite. Many contemporary IQ composites centre on 100 with a standard deviation of 15. About two-thirds of the norming group fall between 85 and 115 if the distribution is close to normal. The scale describes relative standing, not units of intelligence found inside the head.

Norms answer, “Compared with whom?” A ten-year-old and a forty-year-old may earn the same standard score from different raw performances because age changes expected performance. A norm becomes stale if the population shifts while the test remains fixed. A translated test may preserve words yet change difficulty, familiarity or cultural meaning. Responsible adaptation therefore requires more than replacing vocabulary. It requires evidence that the construct, instructions, item behaviour and score interpretations still travel.

Reliability asks how consistently the measurement behaves. Items intended to sample the same broad ability should show enough coherence. Scores should not swing wildly across equivalent forms or short intervals without reason. Yet perfect stability would be suspicious in a human measure. People learn, tire and change.

Validity is the larger question: what evidence supports the interpretation of the score for the proposed use? A battery may predict progress in formal training and still be a poor measure of practical judgement in a chaotic workplace. A subtest may contribute well to a composite but fail when used alone to diagnose a condition. Validity belongs to an interpretation and use, not to a test in the abstract.

Every score has error. A report may provide a confidence interval based on the standard error of measurement. What that interval covers depends on how reliability was studied: task sampling, examiner differences and day-to-day variation are not always included to the same extent. Language mismatch or an unsuitable test can also undermine the interpretation without appearing in the interval. A few points may be noise, especially when several subtests are compared and the largest gap is selected afterwards.

Test difficulty must cover the intended population. A battery with only easy items creates a ceiling among high performers; one with only hard items creates a floor. Modern item-response models estimate how the probability of a response changes across the underlying scale and help developers build shorter or adaptive tests. The model still rests on assumptions about dimensionality, item behaviour and the population used to estimate it.

Standardisation does not require identical treatment when identical treatment blocks access. Enlarged print, signed instructions, breaks or alternative response modes may remove irrelevant barriers. The question is whether an accommodation preserves the construct and whether the available norms still support the interpretation.

Retesting adds another complication. Familiarity with the format, remembered strategies, reduced anxiety and coaching can improve scores.

Meta-analyses find average retest gains, with size depending on interval, form and preparation. The gain is real as a performance change. Whether it represents wider cognitive growth depends on transfer beyond the practised tasks.

A test therefore manufactures a score in the respectable sense that a laboratory manufactures a measurement. The result is constructed through controlled procedure rather than invented at whim. Its value depends on whether every construction choice fits the question being asked.

One Number, Several Abilities

A full cognitive battery rarely produces one result. It produces a hierarchy. At the top sits a broad composite intended to summarise performance across varied tasks. Beneath it sit broad abilities, and beneath those sit narrower skills and individual subtests. The hierarchy exists because neither extreme works. One number loses meaningful differences. A pile of unrelated mini-abilities loses the positive manifold.

Raymond Cattell drew an influential distinction between fluid and crystallised abilities. Fluid reasoning concerns solving unfamiliar problems and identifying relations with limited reliance on specific learned content. Crystallised ability concerns accumulated knowledge and the use of language and concepts acquired through culture and education. The names can mislead. Fluid tests still use learned strategies, and crystallised knowledge reflects the ability and opportunity to learn. The distinction describes patterns, not sealed compartments.

Later hierarchical models added broad domains such as visual processing, auditory processing, short-term working memory, processing speed, long-term retrieval and quantitative knowledge. John Carroll's survey of hundreds of factor-analytic studies helped organise these into three levels: narrow abilities, broad factors and general ability. Contemporary assessment often draws on this Cattell-Horn-Carroll tradition, though tests differ in labels, structure and evidence.

The broad composite is usually the most reliable score because it averages

across more observations. That statistical strength can conceal a practical weakness. If a person cannot hold spoken instructions in mind but reasons well when information remains visible, the composite may predict general learning while missing the route by which ordinary teaching fails. Interpretation moves between levels rather than choosing one forever.

Profiles can be useful when the differences are large, reliable and connected to the referral question. A child with stronger reasoning than rapid output may prompt examination of whether speed demands obstruct access. A person with acquired brain injury may show a change from expected premorbid functioning. A clinician assessing intellectual disability considers both intellectual functioning and adaptive behaviour, including conceptual, social and practical demands, with onset during the developmental period. An IQ near a conventional threshold may prompt attention; it cannot complete the diagnosis.

Profiles are also easy to overread. Give enough subtests and almost everyone will have a highest and lowest score. The gap may reflect measurement error, ordinary variation or a feature common in the norming population. Labels such as “visual learner” can then be built from a thin difference and converted into teaching prescriptions unsupported by evidence. A statistically unusual profile is not automatically clinically meaningful, and a clinically important difficulty may not create a rare score pattern.

Cut-offs deserve the same caution. Gifted programmes, disability services and selection systems need administrative rules, so they often choose thresholds. A line can make access consistent, but nature has not placed a cliff at 70, 100, 120 or 130. Two people on opposite sides of a cut-off may be more alike than two people admitted under the same label. Good systems use thresholds as decision rules while preserving judgement near the boundary.

Multiple-intelligences theory became popular because it protested against reducing human worth to school-like reasoning. Musical, bodily, interpersonal and other talents deserve recognition. The stronger

psychometric claim, that these form many independent intelligences, has not held up well when directly tested. Cognitive domains tend to remain correlated, while some proposed intelligences look closer to personality, interest, skill or achievement.

Emotional intelligence has the same naming problem. Ability measures ask people to identify emotions or reason about how they change. Mixed and self-report measures often ask about empathy, confidence, regulation and social behaviour, overlapping with personality and self-belief. Emotion knowledge can matter without forming one independent rival to cognitive ability. The label covers several measures, so a claim about one cannot be transferred to all.

Savants make the structure vivid in another way. A person can display an exceptional narrow skill, such as calendar calculation or musical reproduction, alongside substantial difficulty in other domains. That disproves a flat ranking of every capacity. It does not erase the correlations found across ordinary batteries or turn each isolated performance into a separate general intelligence.

The number is a summary, not a veto on the detail.

Intelligence Is a Developing System

Francis Galton hoped elementary bodily measures would disclose inherited mental power. At his London laboratory, reaction times, sensory acuity, grip and head dimensions looked satisfyingly physical. They were poor substitutes for the complex performances people meant by intelligence. Modern scanners are more informative, but they create the same temptation: replace an inference from behaviour with a direct reading from the body.

Consider a miniature example, not an intelligence test. Hold the digits 4, 1 and 7 in mind, then put them in ascending order without looking back. You need to retain the digits while doing something to them. Make the sequence longer and there is more to lose during the rearrangement. What looks like one answer depends on storage, attention, a learned rule and control over

the response. A demanding test combines many such operations; one successful answer does not reveal which operation made the difference.

This is why there is no intelligence organ or accepted clinical scan that reads off IQ. Frontal and parietal regions contribute to reasoning and control, but they work with sensory and memory systems through connections spread across the brain. Different tasks recruit these systems differently. Brain-connectivity studies find that several patterns can support similar predictive accuracy. The useful question is how the cooperation works, not which patch should wear the crown.

Total brain volume makes the distinction between association and diagnosis particularly clear. Across large samples, larger volume is modestly associated with higher cognitive scores. The distributions overlap heavily. Knowing a volume does not tell you how someone solved a problem, or whether they did. Age, body size, development and the organisation of the tissue also matter. A bigger container is not an explanation of its contents.

Researchers try to get closer to the operations by measuring processing speed, reaction-time variability and working-memory control. These can help separate components hidden inside complex items. But a reaction-time trial still includes perception, attention, decision and movement. In the digit example, a person might rehearse the sequence, picture it or start sorting before the last digit arrives. The same final score can conceal different ways of working.

Imaging does not remove that ambiguity. More activation can indicate effective recruitment, extra effort or inefficiency, depending on the task and the person's familiarity with it. Machine-learning models can predict some cognitive variation from brain data, but they must work on people whose results were not used to build the model. Predicting a score and explaining the machinery that produced it remain different achievements.

Development keeps changing the machinery. Processing speed and some forms of unfamiliar reasoning often weaken earlier than vocabulary and accumulated knowledge, but there is no single age at which the mind peaks.

Comparing today's younger and older adults also mixes ageing with differences in their schooling, health and experience. Following the same people over time answers a different question.

Such longitudinal studies find meaningful continuity across decades. Childhood scores contain information about later cognitive standing, even as average levels, particular abilities and individual lives change. That coexistence is easy to miss. People can learn a great deal while remaining in a similar order, just as an age-normed score can stay still while the child grows. Stability is evidence of persistence, not a fixed personal ceiling.

Neuroscience matters because it can constrain explanations: damage may separate abilities that usually travel together; repeated scans can track change; experiments can alter the demands on a system. The aim is to explain a developing organism, not replace the person with a scan. Behaviour remains indispensable because learning and solving problems are what the biology has to explain.

Heritability Without Destiny

Ask whether intelligence is genetic or environmental and the grammar has already failed. Genes act through cells, bodies, development and experience. Environments act on organisms with inherited differences. The scientific problem is to explain variation under particular conditions, not to decide which side owns the person.

Twin, adoption and family studies compare resemblance among people with different degrees of genetic relationship. In many studied populations, cognitive performance is substantially heritable, and estimates often rise from childhood into adulthood. The rise need not mean that genes become stronger while environments disappear. People increasingly select, evoke and build experiences that fit their tendencies. A child who finds reading easy may read more, receive different encouragement and accumulate knowledge that makes later learning easier.

Heritability is a ratio of variation within a population at a time. It is not the percentage of one person's intelligence caused by genes. Change the range

of environments, and the ratio may change. A trait can be highly heritable and still respond to intervention. Height remains the standard example because good nutrition changes average height without erasing genetic differences among well-nourished people.

Genome-wide association studies search for statistical relations between measured DNA variants and measured traits across large samples. Intelligence is highly polygenic: many variants contribute effects too small to interpret alone. Polygenic scores combine those associations, but the result is a sample-dependent predictor rather than a reading of genetic potential. Scores built mainly in European-ancestry datasets usually predict less well in other ancestry groups, and accuracy can vary within a broad ancestry label. Population structure, linkage patterns, indirect genetic effects, family environment and measurement all enter the estimate.

Sibling comparisons remove much of the shared family background. Genetic associations can weaken compared with those among unrelated people, yet substantial prediction can remain. The size of the reduction varies with the score, sample and treatment of measurement error. These findings support direct genetic contributions while helping identify indirect family pathways. Here, direct does not mean untouched by experience: a child's inherited tendency to enjoy reading can affect development through the books they read. Genomics is refining the account of development, not making development dispensable.

Environmental effects are not leftovers. Additional schooling raises tested cognitive performance on average in quasi-experimental research, though effect estimates vary by design, age and outcome. Lead exposure is a recognised risk to cognitive development. Severe deprivation, disease, iodine deficiency and interrupted education can produce large losses. More ordinary influences such as school quality, family resources, stress, sleep and health are harder to isolate because they cluster, change over time and affect who encounters which opportunities.

The Flynn effect made environmental responsiveness visible at population scale. During much of the twentieth century, average test scores rose in

many countries, with different gains across abilities and periods. The changes were too fast to be genetic evolution. Schooling, health, nutrition, family size, cognitively demanding environments and familiarity with abstract tests have all been proposed; no single explanation fits every setting. Later cohorts in some countries have shown smaller gains, plateaus or declines. Norwegian comparisons among brothers born in different years indicate that both gains and reversals can arise from environmental changes, but one conscription system is not a universal history.

Group comparisons demand the same level discipline. A mean difference between socially labelled groups describes samples. It does not identify a cause. Race labels combine ancestry with migration, geography, language, schooling, wealth, discrimination, health, neighbourhood and selection into the dataset. They are poor substitutes for precise genetic or environmental variables. Heritability within each group does not reveal why group means differ without additional assumptions.

Present genomic methods do not establish a hereditary ranking of socially defined races by intelligence. That conclusion is about evidential reach, not a declaration that biology never contributes to any human difference or that every observed mean must have one social cause. Current datasets, descriptors and causal designs cannot separate intertwined histories with the confidence such a ranking would require.

Inheritance arrives through an environment, while people partly construct environments through inherited tendencies. The knot is the mechanism. Cutting it into nature and nurture produces two clean pieces that no developing person resembles.

Prediction Becomes Allocation

Intelligence tests survived their history because they predict something. Broad cognitive scores are associated with learning, school achievement, training and performance in many jobs. How much they help depends on what counts as performance and on who is being compared. Predicting an examination result, success in training and a supervisor's overall rating are different jobs for a test.

Prediction is easy to inflate because a correlation describes variation across many people while a decision concerns one. Two distributions can differ in average outcome and overlap heavily. A lower scorer may bring stronger knowledge, persistence or task fit. A higher scorer may be careless, unwell, uninterested or blocked by conditions. The score changes a probability. It does not reveal a future.

Employment research has a built-in sampling problem. Imagine a firm that hires only applicants with high test scores. Among its employees, scores now vary little; the relation with performance may look weaker than it would across all applicants. This is range restriction. Researchers use statistical corrections to estimate the missing relationship, but the correction depends on assumptions about who was excluded and how selection worked.

Some influential older reviews corrected too generously. Reanalyses and newer datasets have produced lower estimates for overall job performance, and recent integrated evidence challenges a simple rule that prediction grows with cognitive job complexity. Other studies still find useful prediction of training and job-specific performance after experience accumulates. The dispute concerns how much, under which conditions, not a choice between infallibility and no relation at all.

Educational and clinical uses have a different hinge. A broad score can help identify unexpected difficulty, estimate prior functioning or contribute to an assessment of intellectual disability. Yet consequences often depend on an administrative threshold. A school with ten specialist places will exclude some pupils who might benefit even if every score is flawless. Scarcity can

masquerade as a property of the child.

Fairness is therefore not one switch. Access asks whether language, disability, cost, equipment or familiarity prevents people showing the intended ability. Measurement asks whether the construct and items function comparably. Prediction asks whether the score-outcome relation travels to the relevant population and setting. Decision fairness asks how errors, benefits and burdens are distributed. These questions can conflict. Equal prediction does not make unequal preparation just, and equal average scores do not guarantee equal consequences.

The alternative to a standardised test is rarely no selection. It may be an interview, teacher judgement, reference, family advocacy, work trial or the ability to wait. These methods can capture qualities a test misses. They also admit accent, status, confidence and resemblance to the selector. Standardisation is valuable because it makes part of the judgement explicit and auditable, not because it removes judgement.

The moral step cannot be outsourced. Evidence that a pupil may struggle can support closer assessment, slower pacing, accommodation or more teaching. It does not by itself justify a reduced curriculum. Evidence that an applicant may learn a technical role faster does not settle how an organisation should weigh current knowledge, safety, teamwork, training cost or wider access to opportunity. Psychometrics estimates probabilities. Institutions choose what those probabilities are allowed to do.

Then allocation bends back towards measurement. Put one child in an enriched class and another in a narrowed track, and the decision changes curriculum, peers, expectations and the material future tests can sample. A diagnosis may bring support, stigma or both. Denying one may leave difficulty interpreted as laziness. Employment selection changes who gains experience and who later appears qualified.

The original performance was real. So is the institution's power to change what happens next. The final question is larger than how well the test predicts. It is what happens after the prediction is believed.

How It Actually Works

The classroom problem

Compulsory primary education in France brought into the same classrooms children who did not progress at the expected pace. Teachers could identify struggle, but their judgements mixed attainment, behaviour, poverty, disability and inconvenience. Physicians working with institutionalised children had classifications of impairment, but those categories did not answer the school's practical question: which pupils might benefit from different instruction?

Alfred Binet and Théodore Simon published a preliminary set of tasks in 1905 inside that problem. Simon was a physician with direct experience of children in care, not a junior footnote to Binet. Together they assembled brief tasks arranged roughly by difficulty: follow a command, name familiar objects, repeat digits, compare weights and define words. The tasks did not imitate the school curriculum closely, but schooling and language still entered the performance.

A full age-graded scale followed in 1908, with a final revision in 1911. Age levels supplied an intuitive comparison: what tasks did children of a given age usually pass? A child performing below the typical level might need closer assessment and support. Binet warned against treating the result as a fixed measure of an unchangeable essence. His programme included exercises intended to improve attention and judgement. The test was a tool for intervention, though the categories around it were never free of the period's assumptions.

That origin matters because the instrument solved two problems at once. It made judgement more consistent, and it made classification portable. A teacher's impression stayed in one classroom. A scale could be copied, revised and attached to administrative rules. The second property would soon dominate the first.

Ranking before testing

Binet did not invent the desire to rank minds. Francis Galton had already tried to turn human difference into a programme of measurement. At his Anthropometric Laboratory in London in the 1880s, visitors submitted to tests of grip, reaction, hearing, vision and sensory discrimination. Galton expected elementary sensory efficiency to reveal inherited intellectual power. The measures proved weak guides to the complex performances people cared about, but the programme supplied an enduring ambition: replace impression with a distribution, then connect the distribution to heredity.

Galton also coined eugenics and treated social position as evidence of inherited ability. The move from measurement to ranking was built into his project. Yet the failure of his sensory tests pushed psychologists towards tasks closer to reasoning, memory and knowledge. Binet's scale succeeded partly because it asked children to do recognisably cognitive work rather than assuming that quick fingers disclosed a superior mind.

The contrast contains the field's first recurring tension. A measure can improve because it moves closer to useful behaviour. The same improvement makes it more consequential when institutions convert the result into a hierarchy of people.

Correlations become a structure

While Binet and Simon were arranging tasks by age, Charles Spearman was studying how performances related. In 1904 he argued that positive correlations among school subjects and tests could be explained by a general factor plus specific components. The mathematical details changed, and modern factor analysis no longer follows his original method, but the conceptual split endured. One tradition built better tasks and norms. The other asked what common pattern the tasks revealed.

Spearman's *g* was not an IQ score. It was a dimension inferred from correlations. Different batteries could estimate it through different tasks, and factor scores depended on the model. This distinction blurred as general

composites became easier to administer than factor theory was to explain. The public received the number; the statistical argument stayed behind the curtain.

Alternative structures appeared almost immediately. L. L. Thurstone emphasised primary mental abilities. Raymond Cattell distinguished fluid reasoning from crystallised knowledge. John Horn expanded the broad factors. John Carroll later surveyed a vast factor-analytic literature and organised narrow, broad and general abilities into a hierarchy. The present settlement is plural but not flat: broad domains matter, and a general pattern remains.

From an age level to IQ

The original age comparison became a quotient through translation and redesign. William Stern proposed dividing mental age by chronological age. Lewis Terman and colleagues at Stanford revised the Binet scale for the United States, published the Stanford-Binet in 1916 and popularised a quotient multiplied by 100. A ten-year-old performing at an age level of twelve received 120. The arithmetic looked like a natural unit.

It worked poorly for adults. Mental age does not keep rising in a simple line, and the meaning of a two-year difference changes across childhood. The quotient also invited reification. Chronological age came from the calendar; mental age was an estimate built from tasks and comparison groups. Dividing one by the other did not make them the same kind of measurement.

Terman believed testing could identify gifted children and improve educational placement. His longitudinal study of high-scoring children helped overturn the stereotype that bright pupils were physically weak or socially doomed. He also accepted eugenic assumptions and wrote about low scores in hereditary and social terms that exceeded what the instruments could show. The same investigator could improve measurement, challenge one prejudice and strengthen another.

The scale becomes a social diagnosis

The American reception changed the meaning of the scale before the quotient was fully settled. Henry Herbert Goddard translated and promoted the Binet-Simon approach while working at the Vineland Training School in New Jersey. In that institutional setting, a tool designed to identify educational need became part of a system for classifying people as normal, deficient or unfit for ordinary social life. Goddard introduced the label *moron* for a supposed higher grade of mental deficiency and argued that the condition was inherited.

His 1912 book about the Kallikak family offered a clean hereditary parable: one ancestral line respectable, another burdened by vice and intellectual defect. The genealogy looked scientific because it included charts, photographs and scores. Later historical work challenged the family reconstruction, the diagnoses and the selection of evidence. The story survived because a pedigree turns tangled lives into one line of descent.

Testing at Ellis Island supplied another stage for overreach. Newly arrived migrants faced unfamiliar language, fatigue and a threatening administrative environment. Selected groups were examined with translated, pantomimed or performance tasks, then results were publicised as evidence of widespread deficiency among nationalities. The samples and conditions could not support a hereditary ranking. Yet the numbers entered a political argument already seeking reasons to restrict immigration.

This was not a case of a neutral instrument suddenly corrupted by one bad user. The construct, sample, label and policy were being built together. Classification changed what counted as evidence about the people classified.

War creates industrial testing

The First World War transformed testing from an individual clinical procedure into mass administration. Under Robert Yerkes, American psychologists developed Army Alpha for literate English-speaking recruits and Army Beta for men who could not read English well. Additional individual examinations were available for some cases. Well over a million men passed through the system.

The practical purpose was classification and assignment. Alpha included analogies, number series, practical judgement and information questions presented in written group form. Beta used mazes, symbol substitution, picture completion and other performance tasks with demonstrations. Neither was cultureless. A picture with a missing part can depend on knowing the pictured object; a maze can reward prior contact with pencil-and-paper puzzles.

The historical importance lay in scale. Instructions, answer sheets, time limits, scoring keys and group rooms turned cognitive measurement into an administrative technology. Examiners assigned letter grades and recommendations within a military hierarchy that needed speed more than a complete psychology of each recruit. Testing acquired institutional prestige because it appeared to sort enormous numbers quickly.

The results also displayed the danger of pretending that a test condition is a transparent window. Recent immigrants, men with limited schooling and men unfamiliar with American language or test conventions performed under circumstances that mixed reasoning with acculturation and education. Yet averages were reported as though they revealed inherited national stocks. Carl Brigham used army data to argue that immigration was lowering American intelligence, then later retreated from the interpretation and acknowledged that the tests could not support clean racial comparisons.

The lesson was available early: standardising the room does not standardise the lives entering it. The lesson did not prevent repetition.

Measurement becomes legal identity

During the eugenic era, labels such as feeble-mindedness joined family history, institutional observation and test results in decisions about segregation and sterilisation. Intelligence testing did not create compulsory sterilisation, and no single score explains the system. It supplied an appearance of technical certainty to political judgements about dependency, sexuality, poverty and heredity.

Carrie Buck's case reached the United States Supreme Court in 1927. Virginia sought to sterilise her under a state law, describing three generations of women as intellectually defective. The evidence was shaped by institutional interests, prejudice and a contrived legal process. The Court upheld the law. The case is remembered for its cruelty, but the mechanism matters: a contested social classification was converted into a medical and legal fact, then used to authorise an irreversible intervention.

Tests also entered immigration, schooling and employment. The injury did not require fraudulent scoring. A measure could be reliable within its procedure and still be embedded in an invalid causal story or an unjust rule. That is why later testing standards came to treat consequences and use as part of professional responsibility rather than as somebody else's problem.

Wechsler rebuilds the adult score

David Wechsler approached adult assessment through clinical work rather than school age levels. The Wechsler-Bellevue scale appeared in 1939. It combined several subtests into a point scale, compared adults with age peers and reported a broader profile, including verbal and performance groupings. Later Wechsler scales became widely used in clinical practice.

Age-referenced standard scores replaced the ratio logic for adult use. A composite could be centred at 100, with spread defined by the norming distribution. The score now meant relative standing among comparable people, not a mental age divided by years lived. This was a conceptual improvement and a quieter description. The public kept the letters IQ.

The batteries also made interpretation more clinical. Vocabulary, similarities, arithmetic, digit span, picture tasks and block designs could reveal uneven performance, though revisions changed the content. An examiner observed how the person approached difficulty, checked sensory and language needs, and read the pattern alongside history. The test was no longer one ladder of questions. It was a structured sample with several routes through it.

That complexity created new temptations. Every profile invited a story. Small subtest gaps were assigned to personality, brain regions or learning styles without enough evidence. Better measurement multiplied the number of ways an interpreter could overreach.

A morning decides a school

After the Second World War, much of England and Wales organised secondary schooling around grammar, technical and secondary modern schools. The eleven-plus helped allocate children between them. It varied by area but commonly combined verbal reasoning, arithmetic, English and related measures. Technical places were scarce, arrangements differed locally and there was never one uniform national examination. A decision presented as a discovery about the child was also a decision about the schools available.

The promise was meritocracy: identify ability rather than purchase advantage. The reality exposed every link in the chain. Coaching and familiarity affected performance. Primary schools differed in preparation. A cut-off converted continuous scores into different institutions. Social class shaped both opportunity before the test and routes after it. Late developers faced a decision made near age eleven. Even where scores predicted later academic performance, the selection rule helped produce the outcome by changing curriculum, peers, expectations and qualifications.

Criticism of the eleven-plus was therefore not one claim. Some attacked measurement error and coaching. Some attacked the belief that ability had settled by eleven. Some accepted prediction but rejected the distribution of

opportunity. These arguments are often confused because they reach the same political conclusion through different links.

Bias enters the centre of the field

Civil-rights disputes in the United States forced test makers and users to answer questions they had often treated as peripheral. Employment procedures that excluded protected groups had to be justified in relation to job performance. Litigation challenged the use of standard IQ tests to place Black children in classes for intellectual disability in California; the resulting restrictions were jurisdiction-specific rather than a national prohibition on intelligence testing. Psychologists studied item bias, differential prediction, construct coverage and the consequences of cut-offs.

The word bias proved too broad to settle anything. An item could have equal statistical difficulty after matching on ability yet still sample experiences distributed unequally. A test could predict the same criterion across groups while the criterion itself reflected unequal opportunity. Removing every item with a group difference could narrow the construct or hide a real educational disparity. Fairness became a family of questions rather than a property stamped on a booklet.

The debates of the late twentieth century also split over group causes. Arthur Jensen argued in 1969 that compensatory education had limits and raised hereditary explanations for Black-white score differences. Later books, including *The Bell Curve*, brought similar claims to a mass audience. Critics challenged genetic inference, test interpretation, samples, social history and policy conclusions. The public argument wanted a verdict. The evidence supplied layers of uncertainty, different causal levels and no ethical instruction for distributing opportunity.

The field writes rules for itself

As testing spread, professional organisations tried to separate technical quality from enthusiasm. Manuals began reporting norming procedures, reliability and relations with external criteria. Joint testing standards developed through the second half of the twentieth century and came to treat validity as an accumulation of evidence for score interpretation and use, rather than a permanent badge earned by one correlation. The 2014 joint Standards codified that approach, making the proposed interpretation and use the centre of the argument.

This changed the questions a competent user had to answer. Does the content represent the intended domain? Do internal relations fit the proposed score structure? Does performance relate to other variables as theory predicts? Do alternative explanations, including language or irrelevant motor demand, account for the result? What consequences follow from the use? The framework did not abolish judgement. It made judgement inspectable.

Psychometric tools became more refined at the same time. Item-response theory modelled the probability of a response at different ability levels, allowing developers to study item difficulty and discrimination and later to build adaptive tests. Differential item functioning asked whether an item produced different response probabilities across groups after conditioning on the ability estimate. These methods can identify a problem; they cannot say on their own whether the source is unfair content, multidimensional ability, unequal learning or a flawed matching model.

The profession's repair was therefore procedural rather than final. State the use, gather relevant evidence, monitor the result and revisit the claim when populations, norms or consequences change.

The modern battery

Current cognitive assessment is built from the accumulated repairs. Developers begin with a construct model, pilot more items than they retain, examine difficulty and discrimination, study reliability, compare the test with other measures and recruit norming samples intended to represent the target population. Translation and adaptation require evidence across languages and cultures. Computerised and adaptive systems can select items based on prior responses, increasing efficiency while adding dependence on software, hardware and model assumptions. Remote testing adds control problems involving identity, distraction, connectivity, equipment and access. A digital interface can reduce some examiner variation while introducing another layer of unequal familiarity.

A clinical assessment begins before the first item. Why was the person referred? What language do they use? Are hearing, vision, motor response, fatigue, pain, medication or anxiety relevant? Which test has suitable norms? Are accommodations needed? The examiner then follows standard rules closely enough to preserve comparison while recording departures that matter.

The result should not be one ceremonial number. It may include a general composite, broad index scores, subtests, behavioural observations, confidence intervals, history, achievement measures and reports of everyday functioning. For intellectual disability, adaptive behaviour and developmental onset are indispensable. For a specific learning question, a cognitive profile alone may not identify the cause. For employment selection, validation must match the job and applicant population.

The chain is more visible than it was in 1905, but it is longer. Better standards can improve each link. They cannot decide what society should do with the result.

How we know

The strongest findings come from several kinds of evidence that answer different questions. Large psychometric datasets establish the positive manifold, score reliability and predictive associations. Factor models test competing descriptions of cognitive structure but do not identify one biological cause. Longitudinal, twin, genomic, educational and imaging studies address development through different assumptions, samples and levels of analysis.

Historical claims rest on original scales, manuals, institutional reports, court records and later scholarship. The record is uneven. Test makers documented procedures more carefully than people classified by them could document consequences. Proprietary modern batteries limit public inspection of items and manuals, partly to protect test security.

Generalisation remains the central weakness. Much genomic and psychological research has used educated, wealthy or European-ancestry samples. Norms age, translations change tasks, and institutional criteria differ. Reported predictive relations depend on how both test and outcome were measured. No single study validates intelligence as one object. Confidence comes from converging patterns, and restraint comes from noticing where the patterns stop converging.

What People Get Wrong

“IQ tests measure how much intelligence you have”

The score invites a substance model. Temperature has degrees, weight has kilograms and intelligence appears to have points. A person with 130 seems to possess thirty more units than one with 100.

IQ points are locations on a normed scale. The test records a pattern of performance, converts raw results through age-based norms and places the composite within a distribution. Change the tasks, reference sample, edition

or conditions and the estimate can change. The score is meaningful because well-built batteries show reliability, coherent structure and useful prediction, not because an IQ point is a natural unit found in the brain.

A score near the ceiling or floor may also be less precise because few suitable items sample that level. A tiny numerical difference can therefore carry less information than its typography suggests.

This correction does not make all scores arbitrary. It changes the grammar. A score estimates relative performance on a defined construct with uncertainty. It should be reported with a confidence interval and interpreted for a purpose. Treat it as a measured quantity of personhood and every later inference becomes easier to overstate.

“There is one intelligence, or there are dozens”

Public debate offers a bad choice. Either *g* is one master power that explains the mind, or each human talent is a separate intelligence and the general factor is a fiction.

Cognitive data support a hierarchy. Different tasks correlate positively, so a broad general dimension is useful. Clusters also remain: acquired knowledge, fluid reasoning, visual-spatial processing, working memory, speed and narrower skills show distinctive variation. Brain injury, development and education can affect them unevenly.

Nor does a general factor require one common biological cause. Developmental reinforcement and overlapping processes can generate broad covariance while leaving several mechanisms underneath the summary.

The many-intelligences reaction became persuasive because a single ranking feels morally and educationally cramped. Musical, bodily and interpersonal talents matter. Calling every valued competence an intelligence, however, does not make the constructs independent or measured as abilities. Some are skills, interests, personality features or achievements.

The hierarchy avoids both reductions. General performance matters, broad abilities matter and specific expertise matters. Which level should govern depends on the question rather than a metaphysical vote for one or many.

“A high IQ means you will succeed”

Intelligence predicts outcomes, which public culture converts into destiny. High scorers are imagined to glide through education and work, while lower scorers are assigned a ceiling.

The evidence supports a probabilistic relation. Cognitive scores are associated with learning, school attainment, training outcomes and performance across many occupations. The strength varies with the task, criterion, sample and statistical corrections. Recent personnel research has revised some celebrated validity estimates downward, while other studies still find useful prediction for job-specific performance and persistence across experience. No single coefficient describes every occupation or outcome.

The overlap remains large. Knowledge, conscientiousness, interest, health, social skill, opportunity, discrimination and luck affect what happens. Success is also a chosen criterion: income, examination marks, scientific discovery, good judgement and a satisfying life are not interchangeable.

Prediction can shrink inside selected groups because universities and employers have already narrowed the range. It can also become partly self-fulfilling when selection into an advanced class supplies harder material, stronger peers and higher expectations.

The correction is less dramatic than either sales pitch. Intelligence changes learning demands and some odds. It does not write a biography, settle human worth or identify who deserves an opportunity.

“Heritable means fixed”

The word sounds like a verdict passed at conception. If intelligence is heritable, education and policy appear cosmetic. If an intervention works, genetic influence appears disproved.

Heritability means something narrower: the share of observed variation associated with genetic differences in a particular population under particular conditions. It can be high where nutrition and schooling are fairly uniform because environmental variation has narrowed. It can change across age, place and time. It says nothing direct about how much one person could change.

Genes also act through development. Children partly select and evoke experiences that fit their tendencies, so inherited differences can become correlated with reading, friendship, teaching and practice. Altering an environment can change outcomes without removing genetic variation.

The same logic appears in inherited medical risk. Genetic influence can be substantial while treatment, prevention and environment still change the outcome. Causation and changeability are separate questions.

Education provides the cleanest correction. Quasi-experimental evidence indicates that additional schooling produces average gains on cognitive tests. The existence of genetic influence does not make those gains impossible. Heritability describes variation in the world as arranged; it is not a forecast for every world that could be arranged.

“Culture-fair tests remove culture”

Non-verbal matrices and picture tasks look cultureless because they avoid vocabulary and factual questions. The hope is understandable: remove language and the test reveals pure reasoning.

No developed mind arrives without learned tools. A matrix requires understanding that symbols form a rule, familiarity with page-based abstraction and willingness to infer the examiner's demand. Pictures assume knowledge of depicted objects. Timed work rewards practice with speeded tasks. Instructions still cross a language and a relationship with the examiner.

Non-verbal measures can reduce specific language and schooling demands. Careful translation, local norms, accessibility changes and

differential-item analysis can expose some mismatches. That is a real improvement. It is not purification.

Accommodation sharpens the point. Enlarged print or signed instructions may remove an irrelevant barrier while preserving the intended reasoning demand. Unlimited time changes a score intended to include speed. Fair access requires knowing which feature belongs to the construct and which merely obstructs it.

The useful question is not whether culture has vanished. It is whether the task, interpretation and use are defensible for this population, language and purpose. Culture can be studied, reduced as a source of irrelevant variance and represented more fairly. It cannot be subtracted from a mind that developed inside it.

“Average group differences reveal innate racial intelligence”

A difference between sample means is easy to calculate and hard to explain. The hereditary claim feels complete because cognitive performance is partly heritable within many populations and some score gaps persist.

The inference does not follow. Heritability within groups does not identify the cause of a difference between group means without further assumptions. Social race categories are historically and politically produced labels, not uniform genetic populations. They combine ancestry with migration, language, schooling, wealth, discrimination, health, neighbourhood and who entered the sample.

Genomic association does not repair the problem automatically. Polygenic scores inherit the discovery sample, trait measure and population structure from which they were built. Prediction commonly weakens across ancestry groups and can vary within them. Within-family analyses reduce some confounding, but they do not reconstruct the causal histories of socially defined racial groups.

Group distributions also overlap. A mean difference cannot classify one

person reliably or show that two people reached the same score through the same developmental route.

Current evidence does not establish a genetic cause for racial rankings of intelligence. That limit should not be turned into the opposite claim that every observed difference has one known social cause. The causal problem is unresolved at the level public argument usually demands.

The political overreach is clearer than the causal answer. A sample difference is converted into an account of individuals, present conditions are treated as inherited limits, and disputed causation becomes permission to withdraw opportunity. None of those steps is supplied by the mean.

“Because tests were abused, they tell us nothing”

Eugenics, segregation and exclusion give this conclusion moral force. A tool used to sterilise, track or stigmatise people appears contaminated beyond repair.

Abuse is evidence about a measure's limits, users and consequences. It does not erase the positive manifold, score reliability, predictive associations or the clinical value of identifying a broad pattern of need. A well-chosen assessment can support accommodation, distinguish intellectual functioning from adaptive behaviour and replace impressions that may carry more prejudice.

Discarding tests also does not discard selection. Schools and employers may fall back on interviews, references, accent, confidence, family advocacy and the selector's intuition. Those sources can add useful information. They are not neutral by default.

Formal testing has one further advantage: it can be audited. Tasks, norms, error, predictive evidence and decision rules can be challenged in ways that a manager's sense of potential often cannot. Proprietary instruments limit complete public inspection, but their claims remain more explicit than an unrecorded impression.

The defensible response is controlled use. Define the question, choose an

appropriate instrument, examine access and group functioning, report uncertainty, combine evidence, monitor consequences and preserve routes of appeal. The history defeats worship. It does not require blindness.

Use It

Follow the inference chain

When a number about a person appears, do not begin by choosing whether to trust it. Reconstruct the route.

What performance was observed? Which task sampled it? How was the raw response converted into a score? Which reference group supplied the comparison? What construct is the score meant to represent? Which outcome does it predict, in which population? What decision is being proposed?

A claim can be strong at one link and weak at the next. A test may rank performance consistently but fail to support a diagnosis. A group average may be measured accurately while its cause remains unknown. A score may predict training speed while the proposed exclusion remains unjustified.

This method prevents two common reactions: accepting the whole chain because the first measurement looks technical, and rejecting the first measurement because the final policy is objectionable. Challenge the link that fails.

Read the uncertainty, not just the score

A printed integer attracts the eye and deletes the error bars. Restore them.

Look for the confidence interval, norming date, age band and conditions. Ask whether the difference being discussed is larger than ordinary measurement error. Check whether several subtests were compared until an impressive gap appeared. A highest and lowest score will exist in almost every profile, even when no rare pattern does.

Then add uncertainty the formal interval may omit. Was the person ill,

sleep-deprived, distressed or working in a second language? Did visual, hearing or motor demands intrude? Was the test repeated recently? Did the examiner depart from standard procedure for a good reason?

Uncertainty does not mean shrugging. It changes the resolution at which the result can be used. A broad difference may be clear while a five-point ranking is not. Good judgement matches the precision of the decision to the precision of the evidence.

Separate prediction from permission

Suppose a test predicts that one applicant will learn a technical role faster. That fact does not decide whether the applicant should receive the job. The organisation might value current knowledge, persistence, safety record, teamwork or the benefit of training a wider pool. It might redesign the work or offer a probationary route.

The same distinction applies in education. Evidence that a pupil may struggle can justify support, a different pace or a closer assessment. It does not automatically justify a reduced curriculum. Evidence that another pupil is ready for more difficult material may justify acceleration without implying greater human worth.

Whenever a predictive claim appears, ask what moral or institutional premise has been added. Scarce places, cost, risk tolerance and ideas about merit often enter silently. Psychometrics can estimate probabilities. Permission comes from a separate argument that must be stated and defended.

Compare the person with the task and setting

Do not ask whether someone is intelligent in the abstract when the practical question concerns a demand.

A role may require rapid mental calculation, sustained attention, verbal explanation, spatial transformation, memory for procedures or judgement under uncertainty. A broad score can estimate learning across demands, but task analysis reveals where knowledge, tools and design matter. A

person who struggles with spoken multi-step instructions may perform well when the sequence remains visible. A slow reader may reason strongly once decoding pressure is removed.

This lens avoids two errors. One is reducing performance to a global trait when a specific barrier can be changed. The other is ignoring general learning demand because a person has one strong skill. Match the assessment to the real task, then ask which demands are necessary and which are habits of the setting.

Ask what changed before calling a change unreal

A rising score may reflect practice, education, improved health, reduced anxiety or broader cognitive development. A falling score may reflect ageing, illness, fatigue, unfamiliar language or a change of norms. “Real intelligence” is not whatever remains after every influence has been dismissed, because development consists partly of those influences.

Use transfer as the test. Does improvement appear only on repeated items, on similar tasks, across broad domains or in everyday learning? How long does it last? Did the comparison group change? Was the same form used? These questions distinguish several kinds of change without pretending that only one deserves the word.

Population shifts need the same care. The Flynn effect shows that average test performance can change across generations. Reversals show that the direction is not guaranteed. Before invoking genetic decline or educational triumph, identify which abilities moved, in which cohorts, under which norms.

Treat profiles as hypotheses, not identities

A profile can direct attention. Weak working-memory performance may suggest checking how a person handles long spoken instructions. Low processing speed may suggest examining motor demands, visual scanning, perfectionism, fatigue or the need for more time. A verbal-spatial difference may matter for a specific learning history.

The profile does not announce its cause. Several mechanisms can produce

the same score pattern, and ordinary people show uneven results. Translate the pattern into a testable question: does the difficulty recur in school, work and daily life? Does an accommodation change performance? Do achievement tests, history and observation agree?

Avoid identity labels built from one battery. “I am a visual thinker” may describe a useful preference or strength, but it should not become a ban on learning through language. “My processing speed is low” should not become a forecast for every timed decision. A profile earns meaning through convergence with life.

The limits

Intelligence testing cannot measure the full range of human capability. It samples structured cognitive performance better than wisdom, moral judgement, creativity in an open domain, courage, kindness, practical knowledge, leadership under pressure or the ability to build trust. Adding those virtues to a questionnaire and calling the result a wider intelligence does not solve the problem.

Scores also inherit their institutions. A technically sound test cannot create enough specialist teachers, remove poverty, make a workplace accessible or decide a fair distribution of opportunity. Good measurement can reveal a difference that policy lacks the will or resources to address. Replacing a cut-off with an algorithm does not remove this limit. It can hide the rule inside a system fewer people are able to question.

Causal claims remain weaker than descriptive ones. Psychometrics can show stable variation and prediction more securely than it can explain every developmental path. Brain imaging, twin designs and genomic scores each add evidence through assumptions and samples. None watches genes, schools, bodies and social treatment separate cleanly across a life.

Finally, no interpretation is harmless merely because it is statistically literate. Labels alter expectations, and expectations alter opportunity. The possibility of feedback belongs inside the use, not in an ethical appendix added afterwards.

The one thing to keep

Imagine a pupil whose assessment shows a broad, recurring learning difficulty. The result is sound: suitable norms, careful administration, uncertainty acknowledged, other evidence in agreement. The question now is what to do on Monday morning.

One school treats the finding as a limit on what the pupil is worth teaching. Another treats it as information about the pace, explanation and support the pupil needs. Both have read the same report. Neither response was printed inside the score.

The second school has no guarantee of a dramatic transformation. Some difficulties persist despite skilled teaching, and pretending otherwise can be another way to abandon the child: promise effortless improvement, then blame someone when it fails. Taking a difficulty seriously means making provision for it, not insisting it must disappear before the person deserves an education.

This is an imagined choice, but it brings the book's history back to the classroom problem. Binet and Simon sought a more disciplined way to recognise need. The portability of their instrument also made it useful to institutions that wanted lasting categories. A measurement could travel farther than the observations that qualified it, until the number arrived alone and looked like a verdict.

You need not choose between refusing to measure and surrendering to the measurement. A test can reveal something important that kindness, confidence or a flattering story would miss. Respecting that finding means understanding what it establishes, not asking it to decide everything else.

The useful question after the report is therefore not whether this is the person's real intelligence, finally caught. It is what this evidence helps us understand, and what we will now make possible or more difficult. The report ends. The pupil goes back to school.

Terms

Intelligence. A broad capacity inferred from learning, inference, unfamiliar problem solving and flexible thought across tasks. It is a construct supported by patterns of performance, not an object directly observed.

Cognitive ability. A capacity involved in mental performance, such as reasoning, memory, speed or acquired knowledge. The term can refer to one narrow skill, a broad domain or general ability.

Latent construct. A trait that cannot be measured directly but is inferred from observable indicators. Intelligence, anxiety and socioeconomic status can all be modelled this way, with different assumptions.

Psychometrics. The science of psychological measurement. It covers test construction, scaling, reliability, validity, norms, item analysis and the evidence needed to interpret scores for particular uses.

Positive manifold. The tendency for scores on varied cognitive tasks to correlate positively. It is the main empirical reason a general cognitive factor and broad composite scores can be estimated.

General factor, g. A statistical dimension capturing variation shared across many cognitive tasks. It is a useful summary and predictor, but its existence does not identify one mental or neural cause.

Factor analysis. A family of methods that explains correlations among observed variables through fewer dimensions. Results depend on the variables sampled, the model chosen and decisions about extraction and rotation.

Factor loading. A coefficient showing how strongly an observed test relates to a factor within a model. A high loading does not prove that the factor is one biological mechanism.

Hierarchical model. A structure with narrow skills below broader abilities and a general factor above them. It preserves both common performance and meaningful differences among cognitive domains.

Fluid reasoning. The ability to identify relations and solve unfamiliar problems with limited dependence on specific learned content. In practice, every test of it still requires instructions, strategies and prior experience.

Crystallised knowledge. Accumulated language, concepts and information acquired through education and experience. It can remain strong when some speeded or novel-problem abilities decline, though trajectories differ among people.

Working memory. The capacity to maintain and manipulate information while carrying out a task. Tests differ in whether they emphasise storage, attention control, sequencing, language or mental transformation.

Processing speed. Efficiency on simple or overlearned cognitive operations, measured under time limits. Scores may also reflect vision, motor output, caution, attention and familiarity with rapid paper or screen work.

IQ. A conventional standard score summarising performance on an intelligence battery. Modern IQ usually indicates relative standing within age-referenced norms, not mental age divided by chronological age.

Norm. A reference distribution built from a specified sample. Norms turn raw performance into relative standing and must fit the examinee's age, population, language, test edition and intended interpretation.

Standard score. A score expressed on a defined mean and spread. IQ composites often use 100 and 15; subtests may use different scales. Comparable numbers require a common metric and suitable norms.

Percentile rank. The percentage of the norming group scoring at or below a result. Percentiles are ranks, not equal units: the score gap between neighbouring percentiles changes across the distribution.

Standard deviation. A measure of spread around a mean. On a common IQ scale, scores of 85 and 115 sit fifteen points, or one standard deviation, from the centre.

Reliability. The consistency or precision of scores across items, forms, raters or occasions. High reliability is necessary for many uses but does not establish that the intended construct was measured validly.

Validity. The evidence supporting an interpretation of scores for a proposed use. It draws on content, internal structure, relations with other variables, response processes and consequences rather than one permanent coefficient.

Standard error of measurement. An estimate of imprecision arising from specified sources of measurement error. It helps form a confidence interval; which sources it covers depends on the reliability study and its assumptions.

Confidence interval. A range around an observed score reflecting measurement uncertainty at a stated level. It discourages false precision, though it remains conditional on the test model and assumptions.

Standardisation. Consistent administration, scoring and interpretation designed to make comparisons meaningful. Appropriate accommodations can preserve standardisation when they remove barriers irrelevant to the construct being assessed.

Measurement invariance. Evidence that a construct and its indicators behave comparably across groups or occasions. Without enough invariance, comparing factor means or score change can mix real difference with changed measurement.

Differential item functioning. A pattern in which groups matched on the measured ability have different probabilities of answering an item correctly. It flags an item for investigation but does not identify the cause alone.

Item response theory. Models linking ability level with the probability of item responses. They help estimate difficulty, discrimination, information and adaptive-test selection, subject to model and population assumptions.

Heritability. The proportion of observed variation in a population statistically associated with genetic differences under current conditions. It is not an individual percentage, a fixed constant or a limit on change.

Polygenic score. A weighted sum of many genetic variants associated with an outcome in a discovery sample. Its prediction depends on ancestry, environment, measurement, sample size and the population where it is applied.

Flynn effect. The historical rise in average intelligence-test performance observed in many populations during parts of the twentieth century. Its size, affected abilities, causes and recent direction differ across countries and cohorts.

Adaptive behaviour. Everyday conceptual, social and practical functioning. Assessment of intellectual disability considers adaptive limitations and developmental onset alongside intellectual performance, because one IQ threshold cannot describe daily support needs.

Go Deeper

The inviting overview

Ian J. Deary, *Intelligence: A Very Short Introduction*, 2nd edition (Oxford University Press, 2020). Deary offers a compact route from scores and the positive manifold to development, ageing, health, genetics and prediction. He writes from the centre of modern differential psychology without treating intelligence as the whole person. Begin here for the mainstream empirical case in an evening. The warning is contained in the format: compression leaves little room for the institutional history and moral argument that make testing controversial. Pair it with one of the historical works below rather than treating a scientific overview as a complete account of use.

The original test

Alfred Binet and Théodore Simon, *The Development of Intelligence in Children: The Binet-Simon Scale*, translated by Elizabeth S. Kite (Williams & Wilkins, 1916). This collection lets you watch tasks and observations become an age-graded scale. Read it for the gap between a practical instrument and the fixed inner quantity later users imagined. The two authors' joint work is visible here, before later retellings pushed Simon into the background. The language of disability is dated and often offensive. That is part of the historical evidence, not terminology to adopt. Read the procedures alongside the authors' cautions about interpretation and education.

The full argument

Nicholas J. Mackintosh, *IQ and Human Intelligence*, 2nd edition (Oxford University Press, 2011). Move here when a short overview no longer feels sufficient. Mackintosh works through test construction, cognitive structure, genetics, environment, group differences and prediction while showing why apparently simple disputes depend on several kinds of evidence. It is more technical than Deary and sometimes expects comfort with correlations and experimental reasoning. The reward is a careful account of what the field can defend without pretending that every controversy has closed. Some genomic evidence is now newer, but the conceptual distinctions and treatment of competing explanations remain useful.

The sceptical history

Stephen Jay Gould, *The Mismeasure of Man*, revised and expanded edition (W. W. Norton, 1996). Gould's attack on biological ranking made the history of craniometry, intelligence testing and reification part of public culture. Read it for the recurring move by which a constructed measure becomes a natural hierarchy and then a political permission. Read it critically too. Specialists have disputed parts of his statistics, selection and treatment of individual researchers, and later psychometrics is not answered by exposing earlier abuse. Its lasting value is adversarial: it forces every confident number to disclose the assumptions and institutions underneath it. The best reading is neither reverence nor dismissal, but comparison with current standards and evidence.

Notes and Sources

The notes identify the evidence carrying the book's main claims and the places where interpretation requires restraint. Current standards, definitions, consequential source claims and recent genomic evidence were checked on 5 September 2026.

Intelligence as a construct. The broad definition, the distinction between observed performance and latent ability, and the conclusion that intelligence is important without exhausting human capability follow Ian Deary's 2012 review, the American Psychological Association task-force report by Ulric Neisser and colleagues, and the later review by Richard Nisbett and colleagues. These sources differ over mechanisms and social interpretation, which is why the body defines the construct through recurring performance rather than one disputed essence.

The positive manifold and general intelligence. Spearman's 1904 paper is the historical source for the general-factor argument. Carroll's 1993 survey is the main basis for the hierarchical treatment of narrow, broad and general abilities. The mutualism model of van der Maas and colleagues and the process-overlap theory of Kovacs and Conway support the distinction

between a robust covariance pattern and an unsettled causal explanation. The manuscript therefore treats *g* as a statistical model with predictive value, not a discovered organ or substance.

Fluid, crystallised and hierarchical abilities. Cattell's 1963 paper is the primary source for the fluid-crystallised distinction. Carroll supplies the three-level synthesis. Labels and exact factor structures differ across batteries, so the body presents the hierarchy as a family of models rather than one universally fixed list.

Test construction, norms and score meaning. The 2014 *Standards for Educational and Psychological Testing* is the principal authority for reliability, validity, fairness, standardisation and responsible score use. The International Test Commission's second-edition adaptation guidelines support the claims about translation, cultural meaning, local evidence and norms. Embretson and Reise supply the account of item-response theory. The common IQ convention of mean 100 and standard deviation 15 applies to many composites, not every subtest metric. Pearson's 2015 WISC-V technical report illustrates conversion from sums of scaled scores to composite scores. The Standards distinguish different sources of measurement error: a reported confidence interval covers those represented by the reliability procedure, not every possible threat to interpretation. The 2014 Standards remained the current completed joint edition when checked; work towards revision does not constitute a replacement standard.

Retesting and preparation. Hausknecht and colleagues analysed coaching and practice effects across 50 studies, 107 samples and 134,436 participants, reporting a corrected overall effect around 0.26 standard deviations with substantial variation. Scharfen, Peters and Holling provide the broader cognitive-ability retest meta-analysis. The body does not assume that a retest gain transfers equally to unpractised tasks.

Intellectual disability and adaptive functioning. The current AAIDD framework, set out by Schalock, Luckasson and Tassé, defines intellectual disability through significant limitations in intellectual functioning and

adaptive behaviour, originating before age 22. Clinical systems differ in wording and classification, but they agree that an IQ threshold alone is insufficient. The body therefore avoids diagnosing from a number and keeps conceptual, social and practical functioning visible.

Multiple intelligences, emotional intelligence and savants. Visser, Ashton and Vernon directly tested several proposed multiple intelligences and found a general factor remained. Waterhouse reviews the evidential weaknesses of independent multiple-intelligence claims. MacCann and colleagues support the narrower idea that ability emotional intelligence can be modelled within a hierarchy of cognitive abilities; Joseph and Newman show that mixed measures overlap with broader traits and methods. Treffert's review supplies the limited savant discussion. These literatures do not justify treating every valued skill as an independent general intelligence.

Brain structure and function. Jung and Haier proposed the parieto-frontal integration theory from converging imaging evidence. Deary, Penke and Johnson review neural correlates while stressing distributed systems and methodological limits. Pietschnig and colleagues estimated a modest meta-analytic association between brain volume and intelligence, around $r = 0.24$, with substantial overlap among individuals. Cox and colleagues provide large UK Biobank evidence on structural correlates. Thiele and colleagues show both the promise and model-dependence of predicting intelligence from brain connectivity. None supports clinical intelligence reading from a scan.

Stability and change across life. Deary's 2014 review synthesises long-term evidence that a meaningful part of rank-order difference persists from childhood into older age. Deary, Pattie and Starr followed 106 members of the Lothian Birth Cohort tested at ages 11 and 90 and reported a correlation of 0.54, or 0.67 after correction for range restriction, on the same broad test. One Scottish cohort does not define every population or age interval. It demonstrates continuity alongside substantial room for change.

Heritability across development. Haworth and colleagues found

increasing heritability estimates from childhood into young adulthood in a large twin consortium. Plomin and Deary review the broader behavioural-genetic evidence, including polygenicity and developmental gene-environment correlation. These are population estimates under studied conditions. They do not allocate a personal percentage to genes or establish that a trait cannot change.

Genome-wide association and polygenic prediction. Savage and colleagues analysed 269,867 individuals and found intelligence-associated loci consistent with a highly polygenic architecture. Mostafavi and colleagues demonstrate variation in prediction within an ancestry group. Howe and colleagues' within-sibship genome-wide analyses found smaller associations for cognitive ability than population analyses, helping separate direct effects from family and demographic pathways. Wolfram and colleagues' 2026 study, reanalysing UK Biobank and ABCD cohorts, reports more limited attenuation for its cognitive polygenic score, including after reliability adjustments, and reduced but substantial prediction in a cross-ancestry comparison. Its editor notes that withheld proprietary algorithm details prevent independent verification of the reported score-variance predictions; public-data benchmark analyses address other findings. The body does not assume one universal size of attenuation or that prediction vanishes outside European-ancestry samples. Nor does within-family prediction settle the causes of between-group means.

Education and environmental harm. Ritchie and Tucker-Drob combined 42 quasi-experimental datasets with more than 600,000 participants. Depending on design and outcome, estimates commonly fell within roughly 1 to 5 IQ points for an additional year of education. The manuscript narrows the claim to average effects on tested cognitive performance rather than a universal change in one mental substance. Lanphear and colleagues' pooled analysis supports the treatment of low-level lead exposure as a risk to children's intellectual function. The examples establish environmental responsiveness, not an exhaustive list of causes.

Secular gains and reversals. Pietschnig and Voracek's meta-analysis

drew on 271 independent samples, close to four million participants and 31 countries, finding large twentieth-century gains with marked variation across domains and periods. Bratsberg and Rogeberg used Norwegian conscription data and within-family comparisons to show that both gains and later reversals could arise from environmental changes across birth cohorts. The Norwegian result is setting-specific and is not presented as a universal direction.

Group differences, race and causal inference. Neisser and colleagues and Nisbett and colleagues review measured group differences and the limits of causal claims. The National Academies' 2023 framework rejects using race as a proxy for human genetic variation. The polygenic studies discussed above concern prediction in particular samples, not the cause of racial group means. Schraiber and Edge show formally why within-group heritability is uninformative about causes of differences among groups without further assumptions. The manuscript's claim is deliberately bounded: sample mean differences can be observed, but current evidence does not establish a hereditary racial ranking of intelligence. It does not claim that every group difference has one social cause.

Prediction of grades, work and social outcomes. Roth and colleagues provide the school-grades meta-analysis. Strenze reviews longitudinal associations with education, occupation and income. Sackett and colleagues re-examine range-restriction corrections and report lower contemporary estimates for overall job performance than several classic reviews. Demeke and colleagues then integrated three overlapping modern meta-analytic datasets, removed duplicate samples and estimated mean operational validity around 0.18 to 0.22 for overall job performance across 212 samples and 109,396 participants. Published in August 2026, that synthesis draws on reviews covering 1990-2022, rather than observations collected in 2026. It concerns overall job performance and does not erase separate evidence for training or job-specific criteria. Berry and colleagues revisit the validity-diversity trade-off, while Hambrick, Burgoyne and Oswald find that general cognitive ability continues to predict job-specific performance across experience levels in a large military dataset. A 2025

erratum corrected the dotted-line adverse-impact ratios in Berry and colleagues' Figure 1 without changing that paper's conclusions. The body presents this as a dispute about magnitude, criteria and correction rather than a choice between universal dominance and zero validity.

Binet and Simon. Binet and Simon's translated writings provide the original tasks, observations and educational concerns. Brysbaert and Nicolas correct textbook myths: Binet was not commissioned by the French Ministry of Education to invent the test, Simon was a substantive collaborator, and the chronology often assigned to the first test is inaccurate. The manuscript therefore describes a preliminary 1905 set of tasks, the first full age-graded scale in 1908 and the final 1911 revision. It also avoids treating Binet as the sole inventor.

Galton and ranking. Galton's *Inquiries into Human Faculty and Its Development* supplies his programme of sensory measurement, heredity and eugenics. Wooldridge and Gould provide contrasting historical interpretations. The body avoids an exact visitor count for the Anthropometric Laboratory and makes only the supported point that elementary sensory measures did not become adequate measures of complex intelligence.

Stern, Terman and the quotient. Terman's 1916 volume documents the Stanford revision and the ratio IQ convention. Mackintosh supplies the historical and psychometric context, including why mental-age ratios work poorly in adulthood. The book separates Terman's contributions to assessment and gifted-child research from the eugenic conclusions he drew.

Goddard, Vineland and the Kallikaks. Goddard's 1912 book is primary evidence for the hereditary parable and its classificatory language. Smith and Wehmeyer's revised history documents the unreliable genealogy, diagnoses and evidential selection behind it. Ellis Island material is retained only at the level supported by the record: selected migrant groups were tested under conditions that mixed language, schooling, fatigue, culture and administrative threat, then publicised beyond what the samples justified.

Army Alpha, Army Beta and Brigham. Yerkes's official 1921 report documents the mass programme, test forms, classifications and examinations of roughly 1.7 million men. Brigham's 1923 book used the results for claims about immigration and national intelligence. His 1930 article later acknowledged that language and cultural differences made simple comparisons of immigrant groups untenable. The body says well over a million rather than treating the administrative count as the central finding.

Eugenics and *Buck v. Bell*. The 1927 Supreme Court decision is the primary legal source. Paul Lombardo's archival history supports the description of a test case shaped by allied institutional actors and a defence that did not protect Buck's interests. The manuscript does not claim that an IQ score alone caused her sterilisation or that intelligence testing created eugenics. It treats testing as one part of a wider system that converted poverty, sexuality, family history and disputed diagnosis into hereditary legal identity.

Wechsler and adult assessment. Wechsler's 1939 book supplies the clinical adult context, point-scale approach and movement away from mental-age ratios. Later Wechsler editions changed the subtests and factor structures. The body uses the historical design principle without presenting any proprietary modern edition as the definition of intelligence.

The eleven-plus. Wooldridge is the main historical synthesis for English intelligence testing and educational selection. Parliament's history of the Education Act 1944 locates the post-war tripartite development in England and Wales. The body keeps local variation, mixed content and the scarcity of technical-school places visible. It makes no single national causal estimate and does not imply that one identical examination operated throughout Britain.

Civil-rights challenges and late twentieth-century controversy. *Griggs v. Duke Power Co.* is the landmark employment-testing decision requiring attention to job relation when practices had disparate racial effects. *Larry P. v. Riles* documents litigation over IQ testing and the placement of Black

pupils in classes for intellectual disability in California. Jensen's 1969 paper and Herrnstein and Murray's 1994 book are cited as influential claims in the hereditary and policy debate, not as final authorities. Neisser and colleagues provide the organised professional review that followed the latter controversy.

Modern batteries and the limits of public inspection. Current practice draws on the joint testing standards, ITC adaptation guidance, instrument manuals and clinical frameworks such as AAIDD's. Proprietary items and manuals restrict open scrutiny partly because repeated public exposure would damage test security. That creates a genuine tension between protecting validity and enabling independent inspection.

Bibliography

Primary and historical sources

Binet, Alfred, and Théodore Simon. *The Development of Intelligence in Children: The Binet-Simon Scale*. Translated by Elizabeth S. Kite. Baltimore: Williams & Wilkins, 1916.

Brigham, Carl C. *A Study of American Intelligence*. Princeton, NJ: Princeton University Press, 1923.

Brigham, Carl C. "Intelligence Tests of Immigrant Groups." *Psychological Review* 37, no. 2 (1930): 158-165.

Buck v. Bell, 274 U.S. 200 (1927).

Galton, Francis. *Inquiries into Human Faculty and Its Development*. London: Macmillan, 1883.

Goddard, Henry H. *The Kallikak Family: A Study in the Heredity of Feeble-Mindedness*. New York: Macmillan, 1912.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Herrnstein, Richard J., and Charles Murray. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press, 1994.

Jensen, Arthur R. "How Much Can We Boost IQ and Scholastic Achievement?" *Harvard Educational Review* 39, no. 1 (1969): 1-123.

Larry P. v. Riles, 793 F.2d 969 (9th Cir. 1984; amended 1986).

Spearman, Charles. "'General Intelligence,' Objectively Determined and Measured." *American Journal of Psychology* 15, no. 2 (1904): 201-292.

Terman, Lewis M. *The Measurement of Intelligence*. Boston: Houghton Mifflin, 1916.

Wechsler, David. *The Measurement of Adult Intelligence*. Baltimore: Williams & Wilkins, 1939.

Yerkes, Robert M., ed. *Psychological Examining in the United States Army*. Memoirs of the National Academy of Sciences, vol. 15. Washington, DC: Government Printing Office, 1921.

Modern research, standards and interpretation

American Association on Intellectual and Developmental Disabilities.

"Defining Criteria for Intellectual Disability." Current criteria page. Consulted 5 September 2026.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, 2014.

Berry, Christopher M., Filip Lievens, Charlene Zhang, and Paul R. Sackett. "Insights from an Updated Personnel Selection Meta-Analytic Matrix: Revisiting General Mental Ability Tests' Role in the Validity-Diversity Trade-Off." *Journal of Applied Psychology* 109, no. 10 (2024): 1611-1634. DOI: 10.1037/apl0001203.

"Correction to 'Insights from an Updated Personnel Selection Meta-Analytic Matrix: Revisiting General Mental Ability Tests' Role in the Validity-Diversity Trade-Off' by Berry et al. (2024)." *Journal of Applied Psychology* 110, no. 9 (2025): 1239. DOI: 10.1037/apl0001308.

Bratsberg, Bernt, and Ole Rogeberg. "Flynn Effect and Its Reversal Are Both Environmentally Caused." *Proceedings of the National Academy of Sciences* 115, no. 26 (2018): 6674-6678.

Brysbaert, Marc, and Serge Nicolas. "Two Persistent Myths About Binet and the Beginnings of Intelligence Tests in Psychology Textbooks." *Collabra: Psychology* 10, no. 1 (2024): 117600. DOI: 10.1525/collabra.117600.

Carroll, John B. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge: Cambridge University Press, 1993.

Cattell, Raymond B. "Theory of Fluid and Crystallized Intelligence: A Critical Experiment." *Journal of Educational Psychology* 54, no. 1 (1963): 1-22.

Cox, Simon R., Stuart J. Ritchie, Mark E. Bastin, et al. "Structural Brain Imaging Correlates of General Intelligence in UK Biobank." *Intelligence* 76 (2019): 101376.

Deary, Ian J. *Intelligence: A Very Short Introduction*. 2nd ed. Oxford: Oxford University Press, 2020.

Deary, Ian J. "Intelligence." *Annual Review of Psychology* 63 (2012): 453-482.

Deary, Ian J. "The Stability of Intelligence From Childhood to Old Age." *Current Directions in Psychological Science* 23, no. 4 (2014): 239-245. DOI: 10.1177/0963721414536905.

Deary, Ian J., Alison Pattie, and John M. Starr. "The Stability of Intelligence From Age 11 to Age 90 Years: The Lothian Birth Cohort of 1921." *Psychological Science* 24, no. 12 (2013): 2361-2368. DOI: 10.1177/0956797613486487.

Deary, Ian J., Lars Penke, and Wendy Johnson. "The Neuroscience of Human Intelligence Differences." *Nature Reviews Neuroscience* 11 (2010): 201-211.

Deary, Ian J., Simon R. Cox, and W. David Hill. "Genetic Variation, Brain, and Intelligence Differences." *Molecular Psychiatry* 27 (2022): 335-353.

Demeke, Saron, Paul R. Sackett, Christopher D. Nye, Yujia Li, and Nathan R. Kuncel. "Is General Mental Ability Still the Best Predictor of Job Performance? Integrating Contemporary Meta-Analytic Evidence." *Journal of Business and Psychology* (2026). Published 25 August 2026. DOI: 10.1007/s10869-026-10146-8.

Embretson, Susan E., and Steven P. Reise. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

Gould, Stephen Jay. *The Mismeasure of Man*. Revised and expanded ed. New York: W. W. Norton, 1996.

Hambrick, David Z., Alexander P. Burgoyne, and Frederick L. Oswald. "The Validity of General Cognitive Ability Predicting Job-Specific Performance Is Stable Across Different Levels of Job Experience." *Journal of Applied Psychology* 109, no. 3 (2024): 437-455. DOI: 10.1037/apl0001150.

Hausknecht, John P., Jane A. Halpert, Nicole T. Di Paolo, and Meghan O. Moriarty Gerrard. "Retesting in Selection: A Meta-Analysis of Coaching and Practice Effects for Tests of Cognitive Ability." *Journal of Applied Psychology* 92, no. 2 (2007): 373-385.

Haworth, Claire M. A., Margaret J. Wright, Michelle Luciano, et al. "The Heritability of General Cognitive Ability Increases Linearly from Childhood to Young Adulthood." *Molecular Psychiatry* 15 (2010): 1112-1120.

Howe, Laurence J., Michel G. Nivard, Tim T. Morris, et al. "Within-Sibship Genome-Wide Association Analyses Decrease Bias in Estimates of Direct Genetic Effects." *Nature Genetics* 54, no. 5 (2022): 581-592. DOI: 10.1038/s41588-022-01062-7.

International Test Commission. "The ITC Guidelines for Translating and Adapting Tests (Second Edition)." *International Journal of Testing* 18, no. 2 (2018): 101-134.

Joseph, Dana L., and Daniel A. Newman. "Emotional Intelligence: An Integrative Meta-Analysis and Cascading Model." *Journal of Applied Psychology* 95, no. 1 (2010): 54-78.

Jung, Rex E., and Richard J. Haier. "The Parieto-Frontal Integration Theory (P-FIT) of Intelligence: Converging Neuroimaging Evidence." *Behavioral and Brain Sciences* 30, no. 2 (2007): 135-154.

Kovacs, Kristof, and Andrew R. A. Conway. "Process Overlap Theory: A Unified Account of the General Factor of Intelligence." *Psychological Inquiry* 27, no. 3 (2016): 151-177.

Lanphear, Bruce P., Richard Hornung, Jane Khoury, et al. "Low-Level Environmental Lead Exposure and Children's Intellectual Function: An International Pooled Analysis." *Environmental Health Perspectives* 113, no. 7 (2005): 894-899.

Lombardo, Paul A. *Three Generations, No Imbeciles: Eugenics, the Supreme Court, and Buck v. Bell*. Updated ed. Baltimore: Johns Hopkins University Press, 2022.

MacCann, Carolyn, Dana L. Joseph, Daniel A. Newman, and Richard D. Roberts. "Emotional Intelligence Is a Second-Stratum Factor of Intelligence: Evidence from Hierarchical and Bifactor Models." *Emotion* 14, no. 2 (2014): 358-374.

Mackintosh, Nicholas J. *IQ and Human Intelligence*. 2nd ed. Oxford: Oxford University Press, 2011.

Mostafavi, Hakhamanesh, Arbel Harpak, Ipsita Agarwal, et al. "Variable Prediction Accuracy of Polygenic Scores within an Ancestry Group." *eLife* 9 (2020): e48376.

National Academies of Sciences, Engineering, and Medicine. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. Washington, DC: National Academies Press, 2023. DOI: 10.17226/26902.

National Council on Measurement in Education. "Testing Standards." Current-edition information page. Consulted 5 September 2026.

Neisser, Ulric, Gwyneth Boodoo, Thomas J. Bouchard Jr., et al.

"Intelligence: Knowns and Unknowns." *American Psychologist* 51, no. 2 (1996): 77-101.

Nisbett, Richard E., Joshua Aronson, Clancy Blair, et al. "Intelligence: New Findings and Theoretical Developments." *American Psychologist* 67, no. 2 (2012): 130-159.

Pietschnig, Jakob, and Martin Voracek. "One Century of Global IQ Gains: A Formal Meta-Analysis of the Flynn Effect (1909-2013)." *Perspectives on Psychological Science* 10, no. 3 (2015): 282-306.

Pietschnig, Jakob, Lars Penke, Jelte M. Wicherts, Michael Zeiler, and Martin Voracek. "Meta-Analysis of Associations Between Human Brain Volume and Intelligence Differences: How Strong Are They and What Do They Mean?" *Neuroscience and Biobehavioral Reviews* 57 (2015): 411-432.

Plomin, Robert, and Ian J. Deary. "Genetics and Intelligence Differences: Five Special Findings." *Molecular Psychiatry* 20 (2015): 98-108.

Raiford, Susan Engi, Lisa Drozdick, Ou Zhang, and Xuechun Zhou. *Wechsler Intelligence Scale for Children-Fifth Edition: Technical Report #1. Expanded Index Scores*. Pearson, 2015.

Ritchie, Stuart J., and Elliot M. Tucker-Drob. "How Much Does Education Improve Intelligence? A Meta-Analysis." *Psychological Science* 29, no. 8 (2018): 1358-1369.

Roth, Bettina, Nicolas Becker, Sara Romeyke, Sarah K. Schäfer, Florian Domnick, and Frank M. Spinath. "Intelligence and School Grades: A Meta-Analysis." *Intelligence* 53 (2015): 118-137.

Sackett, Paul R., Charlene Zhang, Christopher M. Berry, and Filip Lievens. "Revisiting Meta-Analytic Estimates of Validity in Personnel Selection: Addressing Systematic Overcorrection for Restriction of Range." *Journal of Applied Psychology* 107, no. 11 (2022): 2040-2068.

Sackett, Paul R., Saron Demeke, Isaac M. Bazian, Anne Marie Griebie, Reed Priest, and Nathan R. Kuncel. "A Contemporary Look at the

Relationship Between General Cognitive Ability and Job Performance.” *Journal of Applied Psychology* 109, no. 5 (2024): 687-713. DOI: 10.1037/apl0001159.

Savage, Jeanne E., Philip R. Jansen, Sven Stringer, et al. “Genome-Wide Association Meta-Analysis in 269,867 Individuals Identifies New Genetic and Functional Links to Intelligence.” *Nature Genetics* 50 (2018): 912-919.

Schalock, Robert L., Ruth Luckasson, and Marc J. Tassé. *Intellectual Disability: Definition, Diagnosis, Classification, and Systems of Supports*. 12th ed. Washington, DC: American Association on Intellectual and Developmental Disabilities, 2021.

Scharfen, Jana, Judith M. Peters, and Heinz Holling. “Retest Effects in Cognitive Ability Tests: A Meta-Analysis.” *Intelligence* 67 (2018): 44-66.

Schraiber, Joshua G., and Michael D. Edge. “Heritability Within Groups Is Uninformative About Differences Among Groups: Cases from Behavioral, Evolutionary, and Statistical Genetics.” *Proceedings of the National Academy of Sciences* 121 (2024): e2319496121. DOI: 10.1073/pnas.2319496121.

Smith, J. David, and Michael L. Wehmeyer. *Good Blood, Bad Blood: Science, Nature, and the Myth of the Kallikaks*. 2nd ed. Washington, DC: American Association on Intellectual and Developmental Disabilities, 2022.

Strenze, Tarmo. “Intelligence and Socioeconomic Success: A Meta-Analytic Review of Longitudinal Research.” *Intelligence* 35, no. 5 (2007): 401-426.

Thiele, Jonas A., Joshua Faskowitz, Olaf Sporns, and Kirsten Hilger. “Choosing Explanation over Performance: Insights from Machine Learning-Based Prediction of Human Intelligence from Brain Connectivity.” *PNAS Nexus* 3, no. 12 (2024): pgae519. DOI: 10.1093/pnasnexus/pgae519.

Treffert, Darold A. “The Savant Syndrome: An Extraordinary Condition. A Synopsis: Past, Present, Future.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, no. 1522 (2009): 1351-1357.

UK Parliament. "The Education Act of 1944." Living Heritage. Consulted 5 September 2026.

van der Maas, Han L. J., Conor V. Dolan, Raoul P. P. P. Grasman, Jelte M. Wicherts, Hilde M. Huizenga, and Maartje E. J. Raijmakers. "A Dynamical Model of General Intelligence: The Positive Manifold of Intelligence by Mutualism." *Psychological Review* 113, no. 4 (2006): 842-861.

Visser, Beth A., Michael C. Ashton, and Philip A. Vernon. "Beyond g: Putting Multiple Intelligences Theory to the Test." *Intelligence* 34, no. 5 (2006): 487-502.

Waterhouse, Lynn. "Multiple Intelligences, the Mozart Effect, and Emotional Intelligence: A Critical Review." *Educational Psychologist* 41, no. 4 (2006): 207-225.

Wolfram, Tobias, Spencer Moore, Jeremiah H. Li, Jonathan Anomaly, Ivan Davidson, and Michael Christensen. "Interpreting Polygenic Prediction of Cognitive Ability: Evidence for Direct, Reliable, and Portable Genetic Effects." *Intelligence & Cognitive Abilities* 2, no. 1 (2026): 1-19. DOI: 10.65550/001c.158459.

Wooldridge, Adrian. *Measuring the Mind: Education and Psychology in England, c.1860-c.1990*. Cambridge: Cambridge University Press, 1994.