On High-Range Test Construction Now:

Psychometrics, IQ Test Validity, and the Limits of Extreme Intelligence Scores

Scott Douglas Jacobsen

Introduction by Scott Douglas Jacobsen

IN-SIGHT PUBLISHING

On High-Range Test Construction Now: Psychometrics, IQ Test Validity, and the Limits of Extreme Intelligence Scores

IN-SIGHT PUBLISHING
Publisher since 2014

Published and Distributed by:
In-Sight Publishing
Fort Langley, British Columbia, Canada
www.in-sightpublishing.com

Cover Image Credit: Scott Jacobsen on Unsplash

In-Sight Publishing, established in 2014 as a publisher alternative to large commercial publishing houses. Dedicated to the public interest, we remain committed to developing and disseminating innovative projects that are affordably accessible to readers everywhere.

Thank you for supporting independent publishing. Your readership helps sustain a platform that promotes creativity, intellectual freedom, and the amplification of diverse voices.

First edition published in 2026; second edition published in 2026.

Independent Cataloguing-in-Publication Data
Names: Jacobsen, Scott Douglas (author)
Title: *On High-Range Test Construction Now: Psychometrics, IQ Test Validity, and the Limits of Extreme Intelligence Scores*

Author(s): Scott Douglas Jacobsen
Description: Fort Langley, British Columbia: In-Sight Publishing, 2026.
ISBN Canada: Catalogued under In-Sight Publishing.

Identifiers
ISBN (EPUB/Apple Books): 978-1-0673505

Publisher ISBN prefix (In-Sight Publishing): 978-1-0673505; 978-1-0692343

Other In-Sight Publishing titles (selected ISBNs): 978-1-0673505-7-3; 978-1-0673505-6-7; 978-1-0673505-5-0; 978-1-0673505-4-3; 978-1-0673505-3-6 ; 978-1-0673505-2-9 ; 978-1-0673505-1-2, 978-1-0673505-0-5, 978-1-0692343-9-1, 978-1-0692343-8-4, 978-1-0692343-7-7, 978-1-0692343-6-0, 978-1-0692343-5-3, 978-1-0692343-4-6, 978-1-0692343-3-9, 978-1-0692343-2-2, 978-1-0692343-1-5, 978-1-0692343-0-8

Available at: www.in-sightpublishing.com

Disclaimer:
The views and opinions expressed within this publication are solely those of the contributing authors and interviewees. They do not necessarily reflect the official policies, positions, or perspectives of In-Sight Publishing.

Design and Implementation:
Scott Douglas Jacobsen

Layout, typography, font selection, editing, and proofreading:
Scott Douglas Jacobsen

Table of Contents

Acknowledgements

To Alex Tolio, Bob Williams, Chris Cole, (THE LATE) Christopher Harding, Daniel Shea, M.Sc., Dr. Kristóf Kovács, Dr. Ronald K. Hoeflin, Hindemburg Melão Jr., Iakovos Koukas, Marco Ripà, Matthew Scillitani, Patrick Liljegren, Paul Cooijmans, Richard May, Rick Rosner, Roberto Enea, and Seneka, for contributions to this series and permissions for republications from many different times in the history of commentaries on high-range test construction.

Scott Douglas Jacobsen
June 9, 2026

Introduction: Scott Douglas Jacobsen

The Field and Its Intellectual Lineage

High-range test construction remains a niche effort among intelligent and distinctive communities. It is a borderland between puzzle construction and psychometrics, drawing intellectual sport enthusiasts, gifted populations, and, increasingly, artificial intelligence systems. The collection comes from a wide swath of the high-range test-construction communities, with sympathy for the craft and an awareness of its limitations so far. Cognitive differences matter, and intelligence can be studied. Recognition of great ability is important, and serious testing efforts warrant consideration.

The long history of intelligence testing emerged from the Binet-Simon tradition, rooted in the educational needs of certain subpopulations. Spearman's work on general intelligence provided a durable statistical language: a framework for cognitive tasks. Terman's longitudinal work provided developmental studies of high ability. Carroll, Cattell, Flynn, Gardner, Jensen, Sternberg, Thurstone, and Wechsler, among many others, provided an extensive body of work on the true complexity of human intellectual ability. Debate has continued on cultural bias, developmental factors, fluid versus crystallized intelligence, general versus specific abilities, secular IQ gains, test interpretation, and measurement and real-world accomplishment. The collection is a small imprint among the larger impressions of human intelligence made by lifelong scholars of the field.

High-Range Testing Rather Than Ultra-High IQ Testing

The more accurate phrase for this field, to me, seems to be "high-range testing" rather than simply "ultra-high IQ testing." This is an important distinction because some instruments may measure aspects of cognitive ability. Some measure associative range, familiarity, repeated exposure, persistence, spatial imagination, or verbal ingenuity. Many measure an admixture. One set of questions about these communities can be whether a test is beautiful, difficult, or admired. Another set of questions is whether the score assigned to completed test items matches a theoretical intelligence score.

This volume grows out of years of interactions with autodidacts, critics, educators, fraud witnesses, independent theorists, mathematicians, philosophers, poets, physicians, psychometric commentators, society members, technologists, and test constructors. The following is a text about tests and the culture surrounding them. For in-depth individual and group interviews with members of these communities, I would recommend the Some Smart People series through the In-Sight Publishing imprint.

Skepticism, Measurement, and Fragility

Across these conversations, a consistent finding is that high-range testing should be treated with honest skepticism. The field contains genuine ingenuity, historical import, and sincere work. At the highest ranges, most samples begin to thin; therefore, norms become fragile, and confidence intervals widen. Other issues include deterioration in test security in the contemporary era. A

score at the edge of measurement is not a trophy. It should be handled like a volatile chemical: useful in the right conditions, hazardous in theatrical hands.

Mainstream intelligence tests have limits. However, they possess benefits many experimental high-range tests lack: clear interpretive conventions, extensive documentation, large sample sizes, professional administration, and repeated norming. They are useful within practical ranges. The problem at the upper ranges is not that intelligence disappears. The problem is that measurement becomes less secure. A high score may indicate higher ability. It may reflect cultural advantage, norm inflation, exposure effects, technical aids, or test leakage.

Psychometric Guardrails and Fairness

Professional psychometrics offers guardrails: accessibility, administration, documentation, fairness, precision, reliability, and scoring. Reliability does not establish validity. The same is true of difficulty and measurement, or a high ceiling and a meaningful score. Distinctions exist among intelligence, IQ, IQ test scores, and cognitive ability, as well as among achievement, creativity, moral conduct, and wisdom.

A high-range test should be examined for accessibility, construct-irrelevant barriers, differential item functioning, and measurement invariance. A test item may behave differently across groups. The possibility does not make a test useless. Fairness is achieved by testing the conditions under which neutrality fails. High cognitive ability can coexist with disability, executive dysfunction, linguistic differences, and neurodivergence. Accommodations should protect access without changing the construct being measured. The core idea is this: high-range testing must be handled carefully. A score is not meaningful without context. Intelligence cannot necessarily be reduced solely to a single score. A person is larger than a score. Any validity will depend on the test.

Uses, Limits, and Item Quality

Individual interest in the tests comes from entertainment, exploratory ranking, research, self-understanding, and admission to society. Tests should not be used for claims of highest [fill-in-the-blank], overwhelming superiority, clinical diagnosis, educational placement, or employment decisions. Scores should include accommodation status, administration conditions, date, known compromise history, scoring model, and norm and test version. Good test items within these sound testing environments will measure a certain cognitive structure and process, have precision and depth, and require insight. Bad items will have greater ambiguity, brute-force requirements, cultural bias, excessive patience requirements, obscurity, and the inclusion of trivia. Difficulty alone is not a measurement.

Items in these tests tend to be figural, mixed, numerical, spatial, symbolic, and verbal. Each has tradeoffs, with strengths and weaknesses. Language and translation, if verbal, will affect fairness. Boredom and exhaustion can distort results. High-range tests in this regard can reveal unusual forms of reasoning among highly intelligent individuals who fall through the normal cracks of society. However, the tools remain fragile because the field seems young but maturing. Many older tests with older samples have been substantially compromised. Therefore, compromised tests should be retired, restricted, or clearly marked, as is the case with most Mega Society admissions tests.

Ability, Character, and Human Development

Admission to a high-IQ society is an achievement in some measure, but a high score seems almost distinct from creativity, discipline, ethics, or intellectual contribution. In some sense, highly intelligent persons do not have a moral obligation to use their talents, but the world may be better off if they do. Intelligence without purpose and ethics can lead to ego, manipulation, and self-deception, rather than to morally robust character. Perhaps non-cognitive traits determine the tone of intelligence in the music of life.

One big lesson is that if gifted and talented individuals do not receive appropriate support and enrichment, they will seek out intellectual challenges and suitable peers in other domains. Giftedness can bring achievement, confidence, and curiosity, but also alienation, isolation, and self-doubt. A rare score does not confer an easy or stable life. A difficult life does not disprove intelligence. At the end of the day, test takers are people rather than data sets and subject to the same shortcomings as everyone else, except in one regard — having higher general intelligence. Test cultures need clear limitations, consent, ethical restraint, and feedback mechanisms.

High-IQ Societies and the AI-Era Future

High-IQ societies, when done well, are communities with archives, journals, social networks, and admissions systems. They provide a minor sense of fellowship, camaraderie among a cognitive coterie, and a place for unusual minds. At worst, they can foster delusion, credential inflation or fraud, factionalism, and rivalry, but this does not seem to happen more often than in normal life. Darryl Miyaguchi was among the earliest to formalize this commentary, well before any contemporary commentators. Like regular life, the quality and ethics of a community ebb and flow. These communities tend to skew strongly male, and women's underrepresentation should not be seen as a simple map of ability. Many factors sit behind these outcomes. Also, giftedness may be more researched in Europe and North America, while the communities are far from exclusive to those regions.

Finally, in the era of increasingly convincing artificial intelligence, test security becomes more difficult. Authorship becomes even harder. Future testing may have difficulty distinguishing among ability, collaboration, and tool use. Final answers alone may not be enough. To prevent future cases of delusion and fraud among the more theatrical and farcical of the high-scoring claims, the future of test design will likely incorporate adaptive item design, conservative interpretation, correlation with validated tests, independent verification, measurement invariance studies, multilingual construction, privacy-preserving administration, and transparent norms.

Scott Douglas Jacobsen

June 9, 2026

Hindemburg Melão Jr. on the Sigma Test Extended

Hindemburg Melão Jr. is the author of solutions to scientific and mathematical problems that have remained unsolved for decades or centuries, including improvements on works by 5 Nobel laureates, holder of a world record in longest announced checkmate in blindfold simultaneous chess games, registered in the Guinness Book 1998, author of the Sigma Test Extended and founder of some high IQ societies. Melão Jr. discusses: building tests; conclusions about the tests previously; the origin and inspiration for making tests; some definitions and examples of meanings of words; the levels of the Sigma Test Extended; development or improvement of tests; trying to develop questions that tap into a deeper reservoir of skills; the hurdles that candidates tend to have; the process from conception to development and publication; the ideal number of test takers; tests and test builders; and learn from doing this test and its variants.

Scott Douglas Jacobsen: Okie dokie, let's get this show on the road. Like most people in these high-range test construction fields, you are self-taught. A strong point in this is the creativity in testing the construction. When did this interest in building tests really arise for you?

Hindemburg Melão Jr.: First of all, I would like to thank you for the kind invitation to discuss this important subject. It is a topic that has required attention for many years, but has been neglected and even corrupted in recent years. I will comment more on this in response to a related topic.

In 1991, I made drafts of a test I called “Alpha Tests.” Some questions were interesting, but I still had no idea how to create appropriate standards. In 1997, I started accessing the Internet and in 1999 I discovered Miyaguchi’s website, where several high range IQ tests were available. In the same year I founded Sigma Society and reused some of the old questions from the Alpha Tests, along with other new questions, which gave rise to the Sigma Test.

Initially ST was put online in Portuguese, translation software was still very primitive and I am not fluent in English. I tried to do a translation using PowerTranslator 7 from Globalink, but it was very bad. Fortunately, several people became interested in ST and offered to help translate it into other languages, starting with Petri Widsten, who spoke 9 languages fluently. He translated into English, Finnish, French, Italian, and before he began other translations, more people suggested offering to revise details in the Italian and French translations, and to make new translations. In total, it has been translated into a total of 14 languages. In addition to translating, Petri offered the ST for publication in the magazine Mensalainen, from Mensa Finland, and in the magazine IQ Magazine from the International High IQ Society, then Albert Frank published the ST in ComMensal in Belgium and in Gift of Fire by Prometheus. Albert also wrote an article about the ST that was published in Glia’s Papyrus.

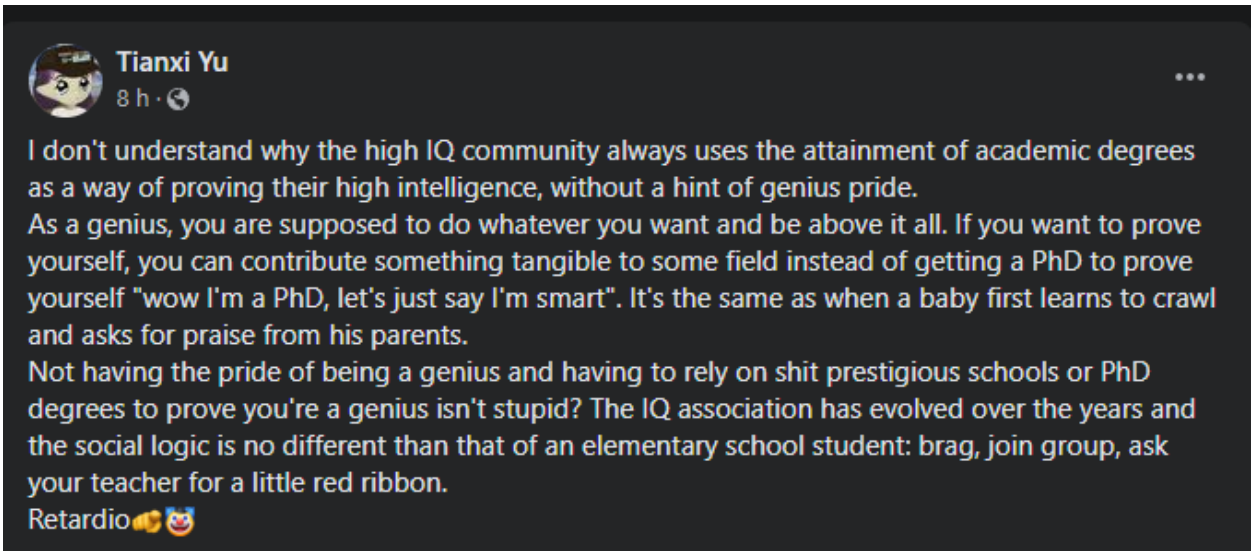
Jacobsen: What were the conclusions about the tests at the time and the need to develop your own?

Melão Jr.: If you don’t mind, I’d rather talk about my impressions of the current tests (which include the oldest ones). I believe my current opinion is more useful.

To begin this response, I would like to analyze two recent comments (a few hours ago and a few minutes ago) posted by Tianxi Yu, in which he touches on important points, which illustrate some of the reasons why I developed new tests, new standardization methods and a new scale.

I started to respond to Tianxi’s message, but soon I exceeded Facebook’s character limit. Furthermore, as I developed the answer, I realized that it would be quite suitable to add as an answer to this question. Considering that the comments are in public posts, I believe that my friend Tianxi will have no objection to them being used here, even because his opinions on this subject are very similar to mine, with few points of divergence. In any case, if he wants me to remove the screenshot, that’s fine with me.

Post 1:



Perhaps what Tianxi meant was not exactly what he said. Some generalizations like “always uses” would not be representations of reality in the context he used. I would almost interpret it as the opposite, and in my network of contacts almost never anyone uses a Ph.D. as “proof” (or corroboration, or evidence) of intelligence. They use it for several other reasons, including because it is an achievement after years of effort in a process of acquiring knowledge and training in the application of the scientific method and certain procedures. They use it for social and intellectual prestige in the eyes of the majority, they use it for commercial, professional, social reasons, etc.

Anyway, I believe that the criticism that Tianxi would like to make, based on the context of what he wrote, is that in general people are more proud of a Ph.D. title than of a corresponding IQ (from 125 , depending on the field and institution) or even a higher IQ, although the title’s rarity level may be lower than the IQ’s rarity level. It would be like a person being proud of some bronze medals in a certain modality than of gold medals in another modality, and this has a derogatory effect on the second modality. In Tianxi’s view, people should feel proud of their genius, and externalize this feeling, and I partially agree with him.

However, people in high IQ societies do not seem engaged in valuing the attributes they have prominently and promoting the recognition of these attributes in the eyes of society. As a result, they lose space to people who “advertise” academic titles that represent less, from an intellectual point of view, but are seen with more admiration and respect by society.

A long analysis would be necessary here, and it would not be possible to analyze all the ramifications. I would select the branch that leads to Andrew Wiles’ criticism of IMOs. Wiles doesn’t place much value on IMO because they are very simple problems that can be solved in 1-2 hours, whereas big real-world problems are much harder and more complex that often take decades or even more than a few generations to be resolved, like what Wiles himself resolved.

There are several points to consider. The first is that IQ is predominantly genetic, the person didn’t have to work hard to achieve it, so I don’t think there would be much reason to be proud. What could be a source of pride is the use of IQ in solving important problems. In this sense, a

typical Ph.D. with 125-135 can contribute more to the common good and the expansion of knowledge than a genius with 190.

This generates discredit and marginalization of high IQ societies, which are not admired or even respected by great intellectuals, nor by the population in general. Most great intellectuals are not even interested in joining these groups. Most of the smartest people are outside of high IQ societies. This does not represent a big problem. But on the other hand, people in high-IQ societies have great potential as "problem solvers", and there are many difficult problems to be solved in the world, but there is no effective connection between these points, resulting in an immense waste of potential. .

I don't want to comment on Kim because I'm irritated by his recent attitudes, and I don't want to run the risk of being unfair with excessively harsh criticism for emotional reasons, but at the same time I can't help but make an objective and impersonal observation about Kim. number cited about Kim's 276 IQ, this is clearly a joke. Most high range IQ tests measure intelligence reasonably well up to about 170, some go as high as 180 but not much beyond that. They may put labels of 250 on the test norm, but the score does not reflect the correct IQ for levels above 180. I have already made attempts to raise this ceiling with the creation of the ST and the STE, but I am aware that I have not succeeded. achieve complete solutions, although perhaps I was able to push the limit a little higher and improve accuracy on the higher scores.

There are truly brilliant people in high IQ societies, but they haven't produced much for different reasons. There are other brilliant people who effectively used their potential in some relevant contributions, such as Petri Widsten, Marco Ripà, João Antonio Locks Justi, Andrew Beckwith. Among those that did not produce, I see some allegations that seem plausible and fair to me, and others that are lame excuses.

I see my own case as an example of a situation of difficulties and many obstacles, my parents were very poor, I live in a backward country where people are prejudiced against intelligence, against science, against logic. I started my degree and stopped after 2 months, so I don't even have half a semester of college left. Despite all this, I improved the works of 6 Nobel laureates in Economics and 1 in Physics and made dozens of original contributions in different fields of knowledge. Objectively comparing my contributions to Economics – especially Econometrics – with those of the winners of the Sveriges Prize for Economic Sciences in Memory of Alfred Nobel, my work is more relevant than that of 90% of the laureates. However, my articles are in Portuguese and are read by few people. Recently, two friends drew attention to this and it is possible that in 2025 I will receive two or more nominations for the "Nobel" in Economics, this depends, in part, on my articles being translated into English and published in indexed journals.

Of course, if I had been educated in a more stimulating environment, I could have produced much more and better, but even in a hostile and impoverished environment, this did not stop me from developing relevant innovations.

Before continuing the argument, I would like to cite one more example: Newton also faced difficulties in childhood and adolescence, according to some authors, Newton cleaned the floors and carried the potties out of his colleagues' rooms, among other similar services, in exchange

for the opportunity to study, but that didn't stop him from achieving extraordinary results. Furthermore, he was beaten by his stepfather and suffered bullying at school, his mother abandoned him as a child to live with a farmer, among other problems. But he was perhaps the person who most expanded the horizons of knowledge in relation to what was understood before and after him.

So there is a bit of unfounded whining from certain people, who could and should produce much more. There are others who cannot be blamed, because they are experts in IQ test questions, they do not produce Science because their specific talent does not encompass the kind of aptitude that real world problems require. So it wouldn't be fair to charge them that. The specific talent for solving questions at the level of difficulty and complexity of IQ tests is like the talent for Chess, or Music, or Mathematics.

In an interview, Fischer said that "he was not a chess genius. He was a genius who played chess, but he could be a genius at any other intellectual activity he chose." This is partially right. He was indeed a genius with multiple talents, but not equally extraordinary. For Chess it was at a level of perhaps 1 in 100 billion. For other areas, such as Mathematics, Physics or Literature, perhaps at the 1 in 1 million level. Therefore he would probably achieve good results in any activity, but not at such a high level as he achieved in Chess.

The best chess players are not necessarily equipped with cognitive faculties for scientific creation at a level similar to what they have for playing chess. In the case of Mathematics, although it involves cognitive processes more similar to those of scientific production, there are still important differences that make it difficult for the majority of great pure mathematicians to achieve exceptional performance in Physics or Investments.

The typical mathematician is excessively concerned with every detail, with extreme rigor and accuracy, while the physicist is content with plausible finite inductions and reasonable evidence. This allows the great physicist to advance quickly in the analysis of very complex problems, while the great mathematician remains trying to demonstrate something in one of the initial stages of the problem and does not advance beyond that point, because for him it is very important to prove each step.

The physicist is satisfied with 99.9% accuracy (or even a little less) or with a sample of 100,000 events, while the mathematician does not accept just 99.99999999999999999999999999999999999% accuracy nor does he accept just googolplexianth corroborating events, or even infinitely many corroborating events (if these infinities do not represent all cases). Ernst Kummer proved Fermat's Last Theorem for infinitely many cases, but it did not represent all cases.

Physicists can build more complex solutions, but with a greater risk of containing errors. In practice, even if there are some "errors", if the approximations are good enough, things end up working. Ptolemy's cosmological model, for example, worked, made reasonably accurate predictions, even though it was fundamentally wrong.

What is useful and sufficient for Physics or Astronomy may not be useful for Mathematics. And if the physicist or engineer tried to achieve the same level of rigor as mathematicians for each

detail, they would spend much more time on each stage and would not be able, in the short span of a lifetime, to produce much of what exists today. Therefore, it would be a naive mistake to believe that a great mathematician would necessarily be a great physicist if he had chosen to study Physics. Certainly a great mathematician is more likely to be a great physicist than a person drawn at random, or even than a person with great ability in another area that requires talents more different from those required for physics than Mathematics. In other words, there is a strong correlation between competence in Physics and Mathematics, but this correlation becomes weaker at higher levels, where specificities become more relevant.

Therefore, Fischer's interpretation was partially correct, he was indeed a genius with multiple talents, but not equally high. At this juncture, the specific ability to solve IQ test questions is not very useful for predicting or diagnosing high intellectual production capacity in the real world, whether in Science or Mathematics. Even IMO problems, which are more like mathematical creation than IQ test problems, are also not good predictors, as Andrew Wiles warned.

That's why one of my main objectives with ST and STE was precisely to fill this gap, creating a test that tries to assess the ability to solve major real-world problems. I was pleased with the result, and the ST and successors (ST-VI, STE, STL) have attracted the attention of some prominent intellectuals, and have received much praise.

Among the people who have done the ST, STE, STL so far, Petri Widsten had 212 and was the author of some innovations and patents, had the best doctoral thesis in Finland in the biennium 2002-2003 for which he also received a Summa Cum Laude distinction, He placed first in some international Logic and Puzzles competitions, including this competition: **http://www.worldiqchallenge.com/rankings.html** . Marco Ripá was 202, he is the author of some innovations in Mathematics and he is still very young, he will probably make other contributions that are even more important than he has done so far. Some people are taking the STE and STL but haven't finished yet, but they are likely to have high scores. Lukas Pöttrich scored above 200 on other tests and at age 8 he scored higher than Terence Tao on the SAT-Math when Tao was 8 years old; Lukas got 800, while Tao got 760, as far as I know, that's a world record. Diego Andrés de Barros Lima Barbosa (Bronze in the World University Mathematics Championship, 1 Silver and 2 Bronze in the Iberoamerican University Mathematics Championship), Federica Zanni (Bronze in IMO) recently registered on the Sigma Society website and spent a long time on the STE page, Kawan Duarte Guimarães Vieira, Davi Filipe de Melo Pereira, João Italo Marques de Lima, José Osmar de Souza Júnior, Mateus Melo and other young talents in Mathematics, Physics, Chemistry, Computer Science, etc. are taking the STL or STE.

It is very gratifying that the ST, STE, STL are also well accepted outside high-IQ societies, being recognized as a psychometric instrument differentiated by its content and standardization methodology. I feel happy and proud about this, because it leads me to assume that there seems to be good agreement about this type of question being suitable for correctly assessing intellectual production capacity in real problems, and people with good experience in solving very difficult problems agree with that. The IMO, despite the limitations pointed out by Wiles, continue to be the best instrument for predicting great talents in Mathematics, and perhaps the

STE is the best for predicting talents for Science, in addition to being the best instrument for intellectual assessment at the highest levels .

I find figure sequence tests interesting because (theoretically) they do not require knowledge, on the other hand they assess a relatively narrow and primitive skill. Chess is heavily saturated with knowledge, but for people who have just learned to move pieces, the kind of skill measured in Chess is better suited to measuring intelligence than the ability to solve series of figures, because in Chess there is much greater complexity and sophistication, in addition to not having a single answer in most cases, but rather a wide variety of answers with different levels of "quality", bearing greater similarity to real-world problems. Even though Chess is more effective, it is still inadequate to correctly assess the intellectual level, especially at the highest levels.

People have broad sets of general skills at a basic level that are strongly correlated, but as progressively higher levels are considered, the skills branch out and capillarize in different ways and the correlations begin to become weaker. In the IQ range of 70 to 140, grades in Mathematics, Physics, Chemistry, Writing, IQ scores generally correlate strongly with each other, between 0.6 to 0.85. But if you consider the range of 140 to 190, the correlation between these same skills becomes much weaker, close to 0.2 to 0.3. A similar effect occurs with IQ tests that use questions that are appropriate to measure correctly in the 70 to 130 range, or the 90 to 150 range, but cease to be appropriate above 160 and even worse above 170, 180, etc.

Another of Tianxi's criticisms that needs to be examined carefully is about people with an IQ of 190 not posting content that he considers compatible with that intellectual level. An exhaustive analysis would take months, but I will try to focus on two points: if a person wants to post photos of cats, or collect license plates (like Sidis), or study alchemy (like Newton) and astrology (like Kepler), this does not reduce her IQ doesn't even cancel out her merits. The person must have freedom to choose their leisure and work activities. But I also understand that if a person exclusively does these things, it can be a waste of potential.

As I mentioned above, the ability to obtain high scores on IQ tests does not imply that the person also has the ability to solve large scientific or mathematical problems. In this case, it is not fair to demand results in Science or any other area. Even if the person has the capacity to produce in Science, I don't think it's right to demand anything from them, but it would be desirable for them to be aware of the importance that their potential represents for the common good, and adopt a compatible stance.

Some people with IQ scores 200+ in IQ tests do not have the necessary attributes for scientific, technological or mathematical production, including high creativity, the ability to maintain focus for years in solving a very difficult problem, the ability to see important details that go unnoticed by the majority, ability to formulate innovative and more effective strategies for solving specific problems that no one had thought of before, etc. YoungHoon Kim is an example, with scores above 200 on some tests, but I know of no evidence that he has solved any really difficult real-world problems.

In the case of Henry Poincaré, when he worked on the 3-body problem, he thought of a completely different approach from what other great mathematicians had been adopting. There

was a huge redundancy between what the entire mathematical community did, as if 1000 mathematicians did almost the same thing. Then Poincaré radically changed the way of analyzing and, in doing so, made important advances. The same when considering Poincaré's work on the shape of the Earth, treating the problem from an unusual perspective and with surprising results, which dramatically expanded our understanding of the subject and even led to the creation of a new branch of Mathematics. Same for Newton, Cantor and others.

High range IQ tests generally do not include questions that adequately assess this type of ability. They just rely on the bet that the kind of skills that work for 90—160 should also work at levels higher than 170, but practical experience has shown that this is not the case. The design of test questions would need to be very different, to require appropriate attributes to measure correctly at the highest levels.

When Leonardo Da Vinci tried to solve the problem of "flying", he did it very differently from what everyone had been doing before him, instead of imitating birds with wings, he tried to understand what was the essence of the physical laws that explained the flight of birds. , and understood that he didn't need wings; could do this with a propeller.

The results achieved by Leonardo show that some important advances do not require decades of work, but rather an insight of a few seconds, although implementation may take months, years, decades or centuries. That's why IMO problems, when the solution depends on this type of insight, end up being more effective in predicting great mathematicians.

In the case of Leonardo's aircraft, the idea was right, but there was no adequate technology, there were no engines with enough power, there were no sufficiently light and resistant materials. There are small flaws in his idea, such as the absence of a second propeller to compensate for the transmission of angular momentum, but he would quickly discover this if he had an engine and light materials that would allow him to test the prototype, and in the first experiments he would detect the errors, correct them and would end up flying. He would not deduce Bernoulli's principle, nor Newtonian dynamics, but he would intuitively understand the relevant phenomena and make the thing work, even without knowing the physical concepts or the underlying mathematical formalism.

Einstein is a very interesting case. In a previous conversation with my friend Iakovos Koukas, he said he thought Einstein wouldn't get 160+ on a modern high range IQ test. I agree, with the caveat that Einstein's correct IQ is well above 200, perhaps around 245 on an interval scale of antilog potentials with mean 100 and standard deviation 16 (obviously the distribution is not Gaussian). This corroborates that IQ tests are not measuring correctly above a certain point. The tests measure anything above 170, but that something is not a faithful and accurate representation of intelligence.

I've already written a lot about this and I won't repeat it here, but in short, clinical IQ tests use questions suitable up to 130. Some tests generate scores of 155, 183, 197 and even more than 200, but the meaning of these scores can only be interpreted as an adequate representation of intelligence up to about 130 on clinical tests and up to 160—170 on most high range IQ tests.

There are two main reasons for this: the difficulty of the questions is inappropriate for higher levels and there is no construct validity at higher levels.

In the article I analyze errors in the WAIS – including psychometric, logical, semantic and epistemological errors – some of the most serious problems I point out are the inadequacy of the tasks to correctly measure up to 155 or 160. Almost all of the sub-tests are very basic. , some of them could be solved by a well-trained chimpanzee. This is useful for evaluating whether an entity (person, animal, AI or ET) can quickly solve tasks with difficulty accessible to an IQ of 80 or less, but solving these tasks very quickly does not indicate an IQ of 100 or 120 or 148.

The psychometric instruments commonly used are good (accurate, reliable, effective) for measuring intellectual capacity up to a certain level. Clinical tests measure up to about 135, regardless of whether nominal ceilings go up to 225, like SB-IV. Some high range IQ tests correctly measure up to around 160 or 170, regardless of whether the nominal scores reach 250.

Some people in high IQ societies have a clear perception of this fact. Others believe (or want to believe) that an IQ of 196 on a test with sequences of figures or numbers is adequate to name one of the 8 most intelligent people alive.

Apparently there is confusion between the meanings of some words, especially the meanings of IQ and IQ test score. Here is an important clarification about the meanings of “IQ”, “intelligence” and “IQ test score”:

Intelligence is an intrinsic ability of the person, which evolves throughout life, generally increasing rapidly until about 15—18 years of age, then continues to increase more slowly until 25—30 years of age, remains almost stable for a few years, and then begins to slowly decline. In my article in which I describe the meanings of the words used in the STL report, I explain this in more detail and present some curves that represent the variation in intellectual level as a function of age.

IQ (intelligence quotient) is the result of mental age divided by chronological age multiplied by 100. If the meaning is changed, the abbreviation must also be changed, replacing the word “quotient”.

Wechsler proposed a different meaning, but continues to use the term “quotient”. An extensive, complex and in-depth discussion would be appropriate here, but I will summarize the main points:

1. On the one hand, as the term “IQ” has become widely known, it would be bad to change it. So let’s preserve the term “IQ”, even if it is not the quotient of a division. However, other important facts cannot be lost sight of: Binet and Simon’s initial idea turned out to be reasonably correct. If the curve of evolution of the intellectual level as a function of age is corrected, instead of using linear growth up to 16 years and stability thereafter, Binet’s idea can be rescued with relative success. There are a few more problems that need to be resolved, but adjusting an appropriate curve is already an important advance. Another point that needs attention is that, in a “panoramic” view over the decades, a smooth curve offers good representation, but in a “microscopic” view over short periods, there are seasonal oscillations in this curve, with seasonality throughout the day, the week of the

year. So although there is growth from 0 to 29 years old, when a person wakes up in the morning , after 7 hours of sleep, at 11 years old, they may be more intelligent than they will be at 12 or 13 after staying awake for 20 hours straight, or with a headache, or under the influence of alcohol. Therefore there are many small fluctuations throughout the day, the week, the year, which can sometimes be greater than the variation in average IQ from one year to another. These short-term fluctuations pose a problem in measurements in supervised testing.

2. A 10-year-old child with the mental age of a typical adult would have an IQ of about 160, but how do we interpret the meaning of this child's IQ when he is a 20-year-old adult? It would not make sense to consider that it would be equivalent to a 32-year-old adult, nor would there be age values in the corrected curve for an adjustment in this case. In this context, the term "IQ" needs a reformulation, as I explain in the "Golden Book of Intelligence".
3. Another important point to consider is that a person who reached the intellectual level of an adult when he was 5 years old is someone who at 5 years old solved problems typical of average adults. This does not mean that this child, when he becomes an adult, will be able to solve much more difficult and more complex problems than an average adult. Generally yes, but not necessarily and not to the same extent. Children like Gauss, Pascal, Galois, von Neumann present, from early childhood, different characteristics that are not present in average adults, and the different attributes of these children are not considered in IQ tests. Children like Ainan Cawley, Adragon de Mello, Michael Kearney, showed abilities of average adults very early, but did not have the differentiated abilities of Gauss or Galois. Sidis's case is at an intermediate level, he had very early abilities of average adults and also had differentiated abilities that are not present in an average adult, although at a level not as notable as that of von Neumann and others.
4. The standard deviation calculated based on IQ measured in this way is about 24 for children (depends on age) and 16 for adults. The standard deviation presents significant variations from one test to another, or one sample to another, but in general it is like this. This provides a physical value for the standard deviation, rather than the almost arbitrary value suggested by Wechsler. What Wechsler did would be like measuring people's heights, finding that there is a standard deviation of 7.23 cm, rounding to 7 cm and changing the entire scale to accommodate that. It is not a recommended procedure and has several undesirable implications. It would only make sense if there was no physical meaning to the standard deviation and the values could be freely manipulated, but that is not the case.

IQ test score is the result of an attempt to measure IQ.

Therefore, there is a person's intrinsic IQ and there is a score that is an attempt to measure intrinsic IQ. People often interpret the score as if it were IQ itself, which is a serious mistake. I've even seen people say that "IQ is the variable measured by IQ tests". It is not. IQ is an inherent attribute of the person, partially genetic, partially influenced by the environment. What the IQ test measures is a set of abilities to perform certain tasks that are assumed to be reasonable

representations of intellectual level, therefore useful for estimating intrinsic IQ. These estimates will be better (more accurate, more reliable) if the questions are more suitable for the level of ability that the test intends to measure.

Considering traditional tests, scores on these tests are usually strongly correlated with true (intrinsic) IQ within a certain range, as long as the test meets certain conditions, especially construct validity for the respective IQ range. Often tests meet conditions in a narrower range than that in which the test is intended to measure, resulting in skewed scores at one or both ends.

This leads to discredit in these scores, because they are not correctly predicting the intellectual level. When Terman selected his 1528 children with IQs above 135 in 1926 and followed the evolution of these children for decades, it became clear that they were in fact much more productive than the population average in cultural, financial, professional and academic success. This is because the tests that Terman used correctly discriminate above 130 and below 130. However, they fail above 130. Two Nobel laureates were examined by Terman and both failed because they were below 130 in the tests applied. Furthermore, there is the famous case of Feynman, who had a score of 123, although he was a Putnam winner, Nobel Prize winner in Physics and author of numerous contributions to Science.

Given this scenario, in order for there to be greater credibility in the results produced by IQ tests at different levels, a broad reformulation of metrics, methods and processes is necessary.

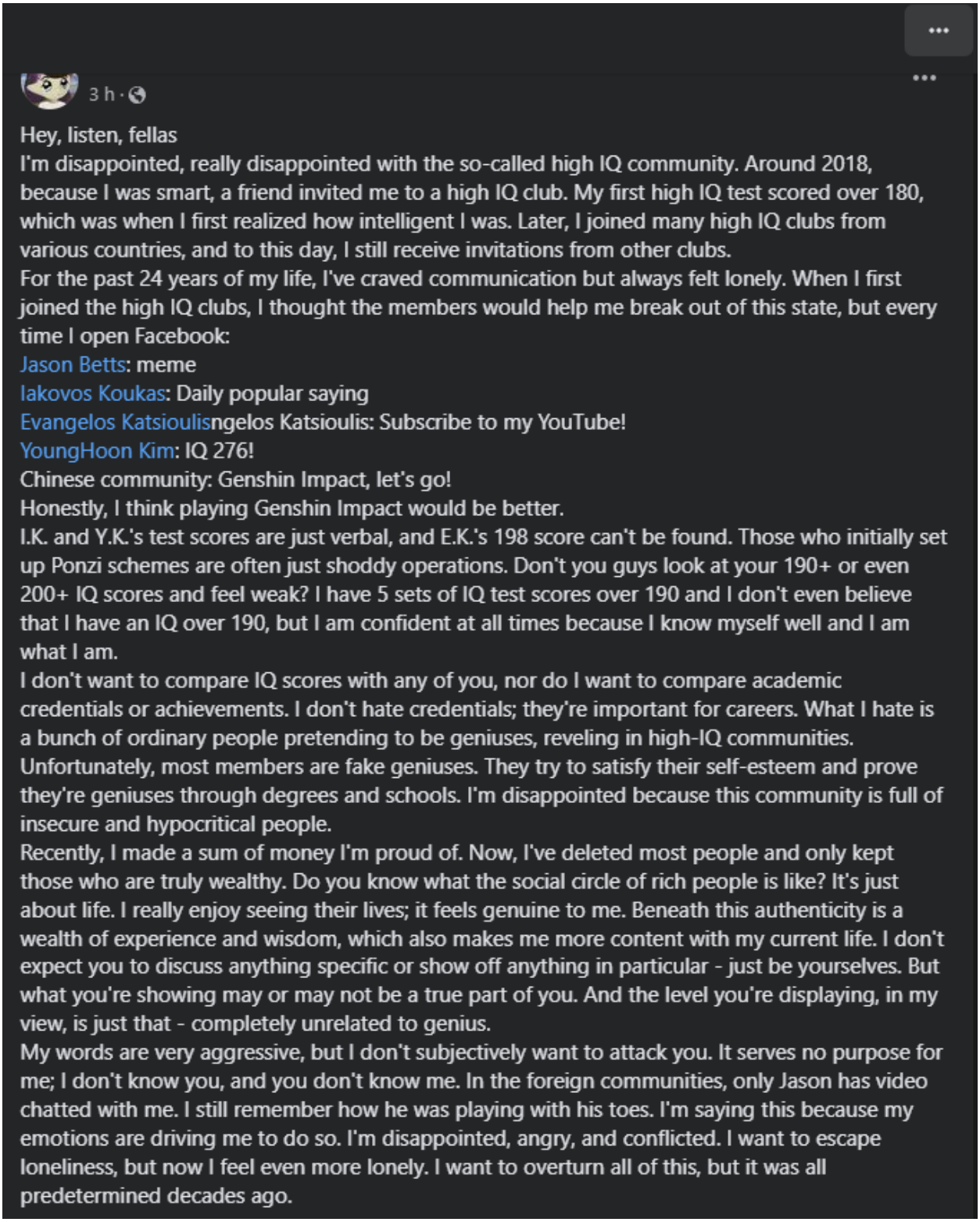
Tianxi talks about “pride of genius”, but what exactly would that be? Proud of finding the next number or figure in a sequence? It might be a difficult sequel, and there’s certainly some merit to that, but it would be better to focus on solving some of the big real-world problems. They don’t need to be BIG, but some problems that broaden the horizons of knowledge and generate benefits for humanity. This seems to me a fairer and more sensible reason to be proud, in addition to being a more correct indication of high intelligence. I am not mixing moral and intellectual criteria in the evaluation process. Creating new and “better” (more effective) weapons, as Archimedes and Leonardo did, are also signs of high intelligence, but applied to the harm of some people. This is part of the thesis I defend. Another part of the same thesis is that it would be desirable to use intelligence for Good, but it is not based on the size of the good generated that intelligence is measured.

I find Tianxi’s point of view interesting, perhaps with small different details. The profile of the person he describes in his critique is perhaps more similar to what is found in some chapters of Mensa. In the case of Mensa Brasil this is common, there are really many people who fit what Tianxi described, but I don’t see many people like that in other high IQ societies. So perhaps the criticism should be directed more precisely at a specific group. Anyway, what I consider important about this are basically 3 items:

1. Correct the bizarre theoretical percentiles, which are obviously wrong in cases far above 130, especially above 160.
2. Improve standardization methods.
3. Improve the content of the questions.

I resolved items 1 and 2 in 2003, item 3 I improved a part in 2000, and continued to improve until 2006, then resumed in 2022.

Post 2:



3 h ·

Hey, listen, fellas
I'm disappointed, really disappointed with the so-called high IQ community. Around 2018, because I was smart, a friend invited me to a high IQ club. My first high IQ test scored over 180, which was when I first realized how intelligent I was. Later, I joined many high IQ clubs from various countries, and to this day, I still receive invitations from other clubs.
For the past 24 years of my life, I've craved communication but always felt lonely. When I first joined the high IQ clubs, I thought the members would help me break out of this state, but every time I open Facebook:
Jason Betts: meme
Iakovos Koukas: Daily popular saying
Evangelos Katsioulisngelos Katsioulis: Subscribe to my YouTube!
YoungHoon Kim: IQ 276!
Chinese community: Genshin Impact, let's go!
Honestly, I think playing Genshin Impact would be better.
I.K. and Y.K.'s test scores are just verbal, and E.K.'s 198 score can't be found. Those who initially set up Ponzi schemes are often just shoddy operations. Don't you guys look at your 190+ or even 200+ IQ scores and feel weak? I have 5 sets of IQ test scores over 190 and I don't even believe that I have an IQ over 190, but I am confident at all times because I know myself well and I am what I am.
I don't want to compare IQ scores with any of you, nor do I want to compare academic credentials or achievements. I don't hate credentials; they're important for careers. What I hate is a bunch of ordinary people pretending to be geniuses, reveling in high-IQ communities. Unfortunately, most members are fake geniuses. They try to satisfy their self-esteem and prove they're geniuses through degrees and schools. I'm disappointed because this community is full of insecure and hypocritical people.
Recently, I made a sum of money I'm proud of. Now, I've deleted most people and only kept those who are truly wealthy. Do you know what the social circle of rich people is like? It's just about life. I really enjoy seeing their lives; it feels genuine to me. Beneath this authenticity is a wealth of experience and wisdom, which also makes me more content with my current life. I don't expect you to discuss anything specific or show off anything in particular - just be yourselves. But what you're showing may or may not be a true part of you. And the level you're displaying, in my view, is just that - completely unrelated to genius.
My words are very aggressive, but I don't subjectively want to attack you. It serves no purpose for me; I don't know you, and you don't know me. In the foreign communities, only Jason has video chatted with me. I still remember how he was playing with his toes. I'm saying this because my emotions are driving me to do so. I'm disappointed, angry, and conflicted. I want to escape loneliness, but now I feel even more lonely. I want to overturn all of this, but it was all predetermined decades ago.

This second post mentions some friends and I prefer not to discuss this point. But generally speaking, I have observed similar problems. In our first In-Sight Journal interview, I already discussed some of these points, so I won't repeat them here. I would just like to elaborate on some previous comments.

ST and STE solve some of the problems that were open, among which the following could be listed:

1. Establishment of a proportion scale. This need was identified by Thurstone in the 1940s and has been the Holy Grail of Psychometrics. Until 2003, the scales were approximately interval for scores below 130 and ordinal when including scores above 130, with distortions in the scale. With my 2003 ST standard I introduced the first scale whose antilogs of scores are on a potential proportion scale, preserving uniform intervals across the spectrum and with a conceptually valid meaning.
2. Improves construct validity, especially at higher levels. Unfortunately, I wasn't able to completely resolve this, but I promoted relevant advances.
3. Adjustment of the difficulty of the questions, seeking to cover all the levels that the test proposes to measure. With STE the real difficulty ceiling of the high range IQ tests rose a few points. Although there may still be, near the ceiling, distortions between nominal and real scores, these are smaller distortions than in other tests.
4. Appropriately weight the points depending on the difficulty of each question. This has several important effects, especially minimizing penalties for carelessness, when a person gets a very difficult question right and gets some very easy ones wrong.
5. Assigning fractions of points to each item, with fair weighting, to refine the score.
6. Review of rarity levels and percentiles associated with each score, especially at the highest levels. I had already written an article about this in 2001 and revised it in 2002, but it was theoretical. In 2003 I gathered data to provide an empirical approach, quantitatively showing the size of the distortions and correcting them. I also calculated new norms for the Mega and Titan, using raw data available on Miyaguchi's website about these tests. The Sigma Test norms were also calculated based on this new methodology, which is explained in more detail in my article **https://www.sigmasociety.net/escalasqi**
7. Determination of the "proportion of potential", as well as the introduction of this concept, which is necessary as part of the standardization process, and also brings some new useful information for different purposes. This is also analyzed in more detail in the article cited above.

In the most recent version of the STE, there were a few more small improvements, including an attempt to determine the curves of variation in intellectual level as a function of age for different IQ ranges. No data from the STE itself was used for this, but rather data on the evolution of the Chess rating as a function of age combined with results from other tests.

At the end of 2023, I started writing the "Golden Book of Intelligence", simultaneously with other books ("Apodictic Guide" and "Project T"). In the "Golden Book of Intelligence" I present

some contributions to Psychometrics, including a review of the WAIS, a review of Richard Lynn's study on the average IQ in several countries, an exhaustive review of the meaning of "intelligence", demystifying some models such as those of Guilford and Gardner, reviewing and improving some concepts such as "fluid" and "crystallized" intelligence, and proposing that the meaning of intelligence varies with IQ, among other topics.

Jacobsen: So, you're the creator of the Sigma Test Extended. You intend this to be the most difficult and reliable cognitive test. What was the origin and inspiration for creating this test – the facts and feelings?

Melão Jr.: I think that in some previous answers I ended up answering this one too. 

Perhaps it is worth commenting a little more on construct validity here, which is extremely important. Several subtests of the WAIS measure latent traits that are not closely related to intelligence, although they are correlated for indirect reasons. This requires a more detailed explanation, and I will use an example to make it more didactic: the "information" subtest has almost no relation to intelligence, they are shallow questions with simplistic answers, they do not require analysis. Despite this, there is a moderate or even strong correlation between intelligence and cultural level, because generally intelligent people also acquire more culture. But this correlation becomes weaker at higher levels and undermines the measurement.

It would be possible to formulate questions that required more complex knowledge, involving analysis. For example: "Why did Einstein, instead of Poincaré or Lorenz, take credit for the Theory of Relativity?" This is the type of knowledge that would lead to a complex and dense discussion, instead of just automatically repeating memorized information, and in this case it would be better related to intelligence, on the other hand, in this example there would be some problems, because the examiner would need to be exceptionally smart and master the topics related to each question. Another problem is that this would be a very specialized question, and if the person being examined did not have much knowledge on the topic, they would not be able to give an adequate answer, even if they were exceptionally intelligent, and in that respect it would be bad.

However, if the test included questions such as those in the WAIS "Information" subtest, it would be desirable for them to be questions that required in-depth and complex analysis, rather than simple repetition and, at the same time, sought to minimize the need for specific knowledge to perform the test. analysis. Even so, there would be the "problem" of requiring exceptional intelligence from the examiner. Therefore, ideally, questions should avoid specialized knowledge, but require thought as part of the answer, rather than simple mnemonic retrieval.

Despite this problem in the "Information" subtest, the scores in this subtest show a moderately strong correlation with the rest of the test and with other tests. This happens because in the range from 80 to 120, generally more intelligent people are also more educated, but above 120, the cultural level progressively ceases to be a good representation for the intellectual level.

We can make an analogy with height, although the correlation between intelligence and height is weaker, the effect is easier to understand. Intelligent people are also generally taller, but it would not be appropriate to include a subtest based on the person's height and include height as part of the total score calculation, because although there is a positive correlation between height and the rest of the test, the correlation weakens at higher levels. higher and becomes practically null above a certain level, generating more spurious noise than contributing to improving measurement accuracy.

If one of the subtests were simply measuring height, a person with an IQ of 2.20 m and 135 on the rest of the test would be no smarter than someone with 1.50 m and 138 on the rest of the test. The same problem occurs when using an "Information" subtest, which impairs measurement at higher levels.

Of course there are some fundamental differences and this analogy is not entirely fair, because culture can provide some tools that help with problem solving, while height cannot (or at least not at the same level). But the point is that the effective weight of culture, of how much culture contributes to the total intellectual level, is much smaller than the weight that the "Information" subtest plays in the total score, resulting in distortions for IQs above a certain level, instead of contributing to making the score more accurate. In other words, high scores on the WAIS would be more accurate if the "Information" subtest, which hinders more than helps, were removed.

In a practical example: a person with an IQ of 150 on the WAIS who got all the Information questions right and got 2 of the Arithmetic questions right is not as intelligent as someone who got all the Arithmetic questions right but 2 Information questions wrong, or even if he got all of them wrong of information. There is a similar problem in the "Vocabulary" subtest, as well as different problems in other subtests.

Jacobsen: What skills and considerations, in general, seem important both for constructing test questions and for creating an effective outline for them?

Melão Jr.: There are several different skills and the lack of some of these skills can be compensated by excellence in others. For example: a vast knowledge of varied issues can compensate for less creativity in creating new issues and vice versa. So there would not be a "closed" set of questions.

Regarding standardization, there are good statistical tools, but cognitive models are still bad. Guilford's opinions add nothing useful and Gardner's opinions bring more problems than solutions. They call these opinions "theories", without any empirical verification or attempt at falsification. In Gardner's case, some recent studies have made it clear that the "multiple intelligences" he proposes are a fantasy. This was predictable and relatively obvious. If Gardner was right, almost every other science would be in trouble using Factor Analysis, which is an important tool in Physics, Astronomy, Economics, Sociology, etc.

The people who promoted relevant advances in Psychometrics were Galton, Cattell (James McKeen Cattell, not Raymond Cattell, whose contributions were minor and unrelated to this specific topic), Pearson, Spearman and Thurstone, in addition to those who contributed to IRT models such as Birnbaum and Lord. I could include Georg Rasch in this list and perhaps a few

others. Binet's works were also important from a different perspective. Wechsler was a disproportionate success, he added half a cent and even made some things worse, in addition to suspicions that I make in my article about WAIS.

The contributions of Pearson, Spearman and Thurstone go beyond the field of Psychometrics and gain space in many other areas. Almost all current major scientific theories use Pearon's linear correlation, Lemaître and Hubble discovered the recession between galaxies using correlation, Henrietta Leavitt discovered the relationship between period and luminosity of Cepheids using correlation, among many other discoveries. Thurstone's contributions were even more notable and could be said to have appeared "ahead of time", only beginning to be more widely used much later, including in AI in recent years and decades.

Analyzing the big names in Psychometrics, the common traits between them, we can intuit some useful characteristics to have a good understanding of the area. In the standardization process, a good understanding of Statistics is important. When preparing questions, it is more difficult to determine what the questions are, as I mentioned in the first paragraph of this answer. But generally creativity and rigorous logical thinking avoid certain problems, as I mentioned in the case of STH at Cooijmans, in our 2022 interview.

Jacobsen: You give some definitions and examples of meanings of words used in the Sigma Test. So any interested reader can get definitions there. Technically, how long has the Sigma Test been in development leading up to the Sigma Test Extended?

Melão Jr.: The first questions that are still present in some Sigma tests were created in 1991, but there was no continuous work throughout that time. In 1991, dedicate a few hours over a few days. In 1999 I dedicated about 1 week to new questions for the ST, with some questions based on known problems and others new ones. The standardization process took longer, and I improved as I received more responses, as with the increase in the number of tests, the use of certain tools and methods that were not possible with smaller samples were being implemented, as well as the creation some new tools and some new methods. In 2007 I closed ST applications.

When the STE was created, I included almost all the questions from the ST and some from the ST-VI, as well as some from the Moon Test. This process took a few weeks. The STL was a joint creation with Tamara, she prepared several questions.

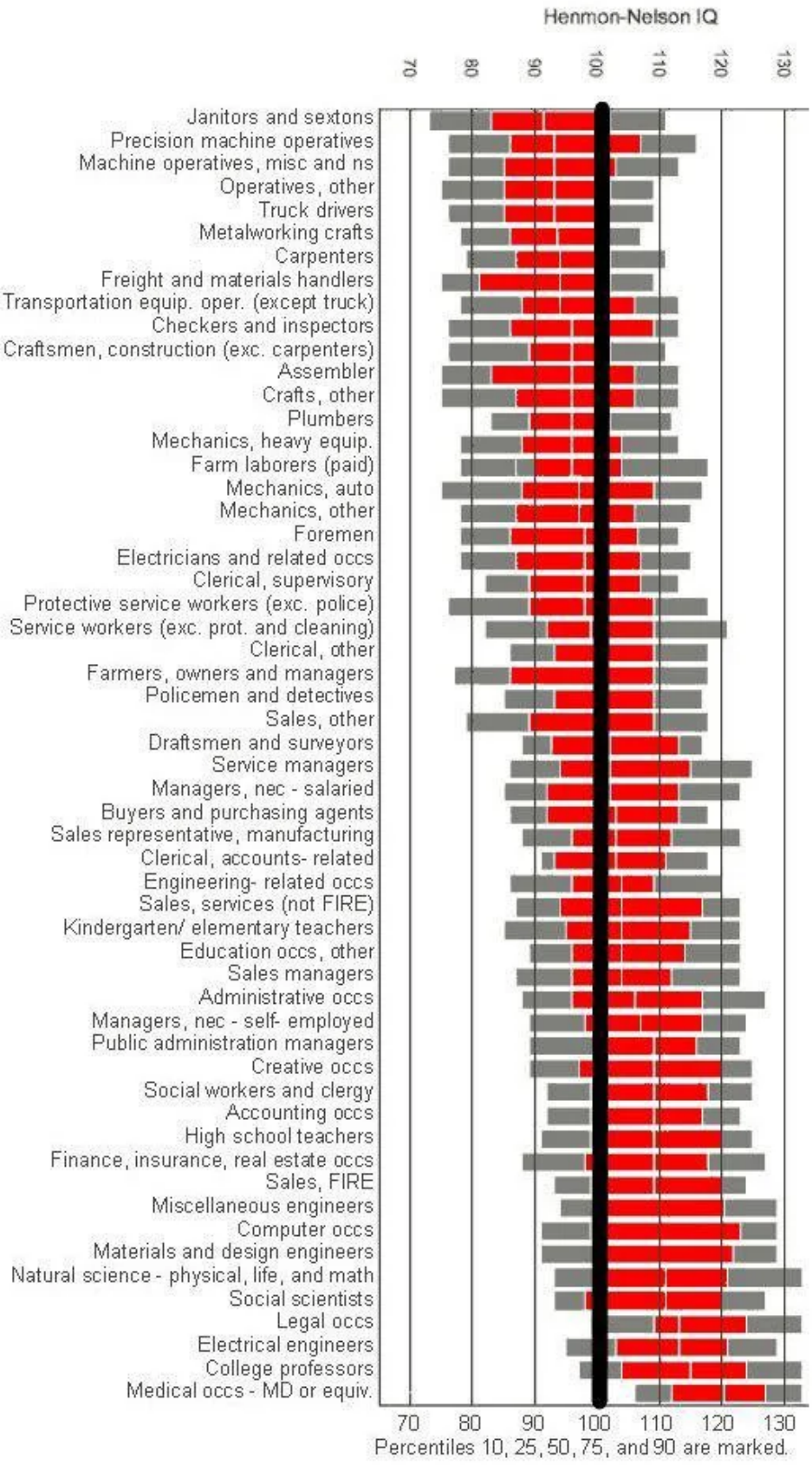
Some differences between the STL and the previous ones are that many questions are in video and photos, showing a real situation from different angles. One can find solutions at different levels and through different methods, just as the methods of Roemer, Bradley, Fizeau, Foucault, Froome and others allow measuring the speed of light with very different strategies, and very different levels of accuracy, the answers can be achieved in different ways. Video questions also make the use of AI difficult, although it is only a matter of time before new AIs emerge.

So the first questions were formulated in 1991, but the total time dedicated to constructing the test was somewhere between 200h and 300h. Time in standardization is difficult to estimate because there have been many updates, but perhaps around 1,000 to 3,000 hours. If you compute the time related to the study and creation of statistical tools, methods, etc., perhaps 10,000 to 30,000 hours, but it would not be correct to interpret this time as applied to this, because many of

the statistical tools developed were for other purposes, especially Econometrics, management risk and genotype ranking.

Jacobsen: You separate the levels of the Sigma Test Extended into Level I (100) Average, Level II (110) Above Average, Level III (120) Superior Intelligence, Level IV (132) Gifted, Talented, High Skills, Level V (144), Level VI (156), Level VII (168), Level VIII (184), Level IX (202), Level X – EXTRA (221). If we correlate these 10 levels to real-world achievements or merit recognition, what jobs, achievements, educational achievements, etc., should we generally expect at each level of the Sigma Test Extended?

Melão Jr.: For scores below 130, it might be useful to reproduce some studies on typical IQs in different professions. Searching on Google, you can find many other lists, tables and graphs like this:



It is important to highlight that in each profession there are quite wide ranges that intersect. We must also remember that Langan, Rosner, Grady Towers have already worked in activities that are very incompatible with their intellectual level, just like me and my father. Therefore, factors such as network and cultural aspects in certain countries may be more relevant than IQ in positioning a person professionally or even academically.

It is also important to remember that specific skills weakly correlated with IQ can play a central role in success and diverse achievements. Nakamura, for example, may not have an IQ above 120 or 130, but he has a very developed talent for chess and has achieved a rating that normally people with an IQ of 180 or 200 may not reach even if they train a lot for it. The same goes for different professions, which may require some specific skills, such as surgeon, where fine motor coordination could not be replaced by any IQ score.

Having made these reservations, we can try to make some estimates of typical achievements for each IQ range.

In this study I review typical IQs at different universities in the USA: **https://www.sigmasociety.net/artigo/qi-universidade-escolas**

With an IQ above 160, depending on the area of activity and the nature of the research carried out, the possibility of winning a Nobel Prize becomes plausible. Although there are cases of Nobel Prize winners with an IQ below 140 and even below 130, what is observed is that the vast majority with an IQ above 160 do not win the Nobel Prize, therefore having an IQ of 160 cannot be interpreted as a predictor of a high probability of a Nobel Prize. , but it can be interpreted as "meeting a minimum requirement" for this. It's not easy to answer this, because exams like the SAT and GRE are not appropriate for testing above 130, and most Nobel laureates have never taken an IQ test with appropriate difficulty and construct validity at their level. Studies that indicate around 155 for the average IQ of Nobel laureates in Science simply reflect the inadequacy of IQ tests to measure at the highest levels. It would be naive to think that Nobel laureates are at the 1 in 3,000 level of intellectual rarity. The most reasonable interpretation is that they were examined with inadequate tests.

A more realistic estimate would be about 170—180 for the average Nobel Science winner, and perhaps 160 is an "inclusive" cut-off point.

In general, most presidents of different countries have an IQ between 120 and 155, rarely above 160 or below 120. Information has already circulated on the Internet that George Bush Sr. would have an IQ of 91 or 102, but he obtained a score of 1206 on the pre-1974 SAT, that would correspond to about 132, which is more plausible for a president with the minimum attributes for his role. Netanyahu is cited as having a 180, I never got around to researching in depth the accuracy of this information and adequacy of this score (the information may be legitimate, but the score may be based on inadequate testing). I think it's reasonable that Netanyahu could actually have something between 160 and 180, but it's a rare case.

Therefore 130 may already be enough to be president in most countries, which represents a serious problem. The problems a president must deal with are extremely difficult and complex, to the point that not even 190 or 200 would be enough to adequately resolve most issues. The big

mistake is that heads of state are appointed based on elections. There should be a better set of criteria, based on the country's effective ability to deal with problems. When David Ben-Gurion invited Einstein to be president of Israel, it seemed to me an extremely intelligent and appropriate invitation, although the methodology (invitation) is very dangerous, it can work if the person (or committee) making the invitation is suitable and competent.

To work at Big Techs, 150 to 160 is usually enough. Champions in IMO and similar generally have around 170 to 190, occasionally they can have much more, but they rarely have much less than 170. Around 170 in conjunction with a lot of training and specific talent for Mathematics or Physics can represent good chances of medals in IMO and other intellectual olympiads. The correlation of IQ with Chess is weaker than with Mathematics, and this correlation decreases at higher levels , so it would not be possible to make many predictions about Chess achievements based on IQ.

People like Musk, Gates, Zuckerberg, Bezos generally have an IQ between 150 and 160, but very few people with an IQ between 150 and 160 reach the level of financial success they did because it depends much more on other factors, including luck, network, discipline , dedication etc. In Leonard Mlodinow's book "The Drunkard's Walk", the author analyzes several cases in which in large population samples the factor of luck can play a large role in determining success at a very high level, and he attributes to Gates and others a great luck. In my opinion, in these cases luck also accounts for most of the result obtained, but talent was also fundamental. If Gates had just been lucky, he obviously would not have developed the products or managed the various situations successfully. Factors related to personality also end up being very important. IQ is just one of the variables in determining economic success, and the weight of IQ depends on several other factors. In some cases IQ can be decisive, in others it can be almost irrelevant.

The cases of Musk and Jobs are a little different. Musk may have an IQ of less than 160, but he appears to be very creative, at a level equivalent to about 180. Jobs scored 1440 on the GRE, which corresponds to about 148, but most likely the GRE did not correctly reflect his IQ, nor creativity, which would be much higher, perhaps at a level of creativity a little below that of Musk.

For awards such as the Fields Medal, Abel Prize, Einstein Prize, the "necessary" IQ is similar to the "necessary" for the Nobel Prize, but accompanied by a set of specific aptitudes for Mathematics. This does not mean that the average IQ of the winners of these awards is similar to the average of the Nobel Prize winners in Science. As the rarity is greater and the questions are similar, I estimate that the average IQ is slightly higher among winners of these awards in Mathematics.

Jacobsen: What does this dimensioning propose as development or improvement of tests like WAIS?

Melão Jr.: The article I wrote about WAIS points out some problems, but from a strictly technical point of view, I don't believe it is appropriate to "fix" WAIS. Due to the number of corrections, it would be more interesting to start something from scratch. However, from a

commercial point of view, as WAIS already has good acceptance, for this reason a broad review could be justified (commercially).

Jacobsen: When trying to develop questions that tap into a deeper reservoir of skills, what is important about verbal, numerical, spatial, and other types of questions?

Melão Jr.: In some cases, it may be interesting to exclusively create sequences of numbers or figures or both. In other cases, tests with analogies and/or associations. In other cases, a diversified test and a heavy dose of randomness in that diversification may be preferable. In the introductory text of the STE I discuss some negative aspects of a test consisting exclusively of sequences of figures, or exclusively of associations, or exclusively of analogies, which results in a very high "internal consistency", and the meaning of this may be a narrowing of skills measurements, redundancy, and other undesirable effects.

The term "internal consistency" should not be the term used. Cronbach's Alpha measures homogeneity, which should not be interpreted as "internal consistency". A very high Cronbach's Alpha indicates that the test measures a very narrow and redundant range of latent traits, and this may not be very useful if the primary goal is to measure the g-factor, which would be a broadly applicable trait.

On the other hand, it has been verified that tests consisting exclusively of sequences of figures, such as Raven, Cattell or some subtests of the WAIS and DAT, present a sufficiently strong correlation with the score in more comprehensive tests, in order to allow the scores in these tests (of figures) are accurate estimates of g at least in the 75 to 125 IQ range and perhaps a slightly wider spectrum.

At levels above 140 and, especially, above 150, the use of these questions becomes increasingly inappropriate. The complexity and difficulty that can be achieved in a test based on a sequence of figures is limited, and they are also solved by exhaustive attempts rather than by brilliant and profound ideas. So what is being measured is something more akin to persistence, patience, determination than intelligence. Some Power Test questions can be resolved in very laborious, time-consuming and non-creative or ingenious ways. The STE also presents this problem in some questions, unfortunately I was not able to completely eliminate this, but in the STE this ends up being a contamination of the question, not the essence of the question, that is, the main difficulty of the question lies in having some creative idea, but part of the solution also requires a laborious and time-consuming process, so I consider it "tolerable", but if the problem can be solved exclusively through a laborious and time-consuming process, without the creative idea, the purpose is lost. In some cases, it is very difficult to avoid the solution being laborious and time-consuming, but one should try, whenever possible, to require creativity and deep thinking in the most difficult issues.

Jacobsen: What are the hurdles that candidates tend to make in terms of thought processes and assumptions about time commitments on these tests? So they get artificially low scores on high-stakes tests.

Melão Jr.: This is an interesting and difficult problem to solve. Perhaps there is no complete solution, because to serve people who do not have a lot of time, it would be necessary to press on

time and harm those interested in engaging in very difficult and time-consuming issues. Andrew Wiles criticized the IMO precisely because the time available is too short (3 hours) to propose challenges with an appropriate level of difficulty and complexity, compromising the purpose of trying to identify future great mathematicians. On the other hand, there would be many operational difficulties if the IMO race took much longer, there would even be the problem of lack of supervision, or the need to host people from several countries for a long time at the competition headquarters, and monitoring them continuously could generate problems related to privacy, since people would need to be supervised after they knew the statement, so if the person took 10 days to resolve an issue, they would need to be monitored so as not to receive help or use prohibited means. Alternatively, supervision could be dispensed with if the issues were unresolved real-world problems.

It would be an interesting idea to hold math and science Olympiads lasting a few months, using much more difficult problems, including unsolved real-world problems, gathering sponsors, etc. But apparently the organizers of these events are satisfied with the way things are.

My focus has been on the correct measures at the highest levels, so I have not been as concerned about the problem you described in this question, but it does represent a source of distortion in scores. On the other hand, I believe that most traditional tests used in clinics already meet this requirement reasonably well, measuring with good precision and accuracy in the range of 70 to 130. I believe that the IQ range in which errors are still large, and they need greater attention, whether at the highest levels, and in these cases time does not seem to be such a demotivating factor, because they are generally much more competitive people and for them it is important to achieve as much as they can, reducing the risk of associated distortions to the time required for resolution.

I also read the text you sent me with the interview with AntJuan Finch and it seems to me that he is already doing excellent work in this regard, as well as Chris Cole, increasing reliability in unsupervised online tests, and encouraging more people to take the tests in a short time and at no cost. With this, I believe that an alternative to clinical tests has emerged with a comparable (or higher) level of accuracy and reliability, accessible to a greater number of people.

Jacobsen: Without spoiling the mental sport of HRTs, what was the process from conception to development and publication of the Levels I to IV STE questions? What was the process from conception to development and publication of the STE questions for levels V through VIII? What was the process from conception to development to publication of the STE questions for levels IX and X?

Melão Jr.: I will try to give an answer by grouping this question and the following two, choosing some items that I consider most interesting to be analyzed individually and making some general comments about all the items.

Some questions are trivial and there would be no way to get away from that much, due to the relatively low difficulty, but even among the questions for levels I to IV I tried to require the person to understand some facts, rather than just applying a formula. I couldn’t go too far into the

explanation without providing some important “clue,” but I can say that some Ph.D.s. in Physics, Engineering and Mathematics missed fundamental details in some questions that seem trivial.

The information that the questions are roughly ordered by difficulty is useful to know that some questions that seem easy are actually not, and there are “hidden” details to be discovered. It’s not a “prank”, that’s not the objective. These hidden details are “natural” and important ones that people should consider but often don’t realize. In some ways they are similar to the Monty Hall problem, which seems simple and obvious at first glance, but when you start to dig deeper you realize that there are subtleties and complexities.

Question 22 is an interesting example that the vast majority got wrong, including astronomers and mathematicians. I even thought about changing the position of this question to a higher level, because if you consider the number of correct answers out of the total number of respondents, it has a lower correct answer rate than questions that are at higher levels. However, I decided to keep it where it is because it is not actually “more difficult”, the problem is that people underestimate the difficulty. There are people from Giga Society who made mistakes, but I believe that if they had “respected” the difficulty more and believed that it was at a level compatible with its difficulty, they would have analyzed it more carefully and would have gotten it right. This comment is in a way a useful “clue”, but I don’t see a problem in providing this clue because the position of this question at level V is also a clue, however people don’t believe it has level V and this leads to error, so I see no harm in reinforcing that “she is really level V and maybe a little higher”.

Question 35 raised a long debate with Peter David Bentley, D. Phil. (=Ph.D.) and Post Doctoral in Physics from the University of Oxford. Petri Widsten and Albert Frank entered the debate. When a person has a score above 180, they are notified of a question they got wrong and they can debate whether they consider their answer should be accepted, and that happened in this case. It was an analysis that lasted several days. (this question was part of the ST, Peter did not take the STE)

Question 50 has a detail that perhaps I should make more explicit, because some people have consulted the distance from the Moon to the Earth in ephemeris software, and this really does not violate the general statement of the test that allows using any available resource. So perhaps I should make it clearer that for this specific question the person needs to use the data available in the photo and text of the statement, which is why higher resolution photos are available for download. When the person resolves it using ephemeris software, I ask them to send it again using the photos.

Question 45 has also received responses in which the person underestimates the difficulty and I ask them to send it again.

People generally realize that there are hidden subtleties that make the problem more difficult than it seems at first glance, but in some items most people don’t notice.

In question 48, I wanted to get an idea about whether people in high IQ societies were aware that the percentiles in groups above 130 are wrong and the error grows at higher levels, as well as I

would like to know if they have a approximate idea of the magnitude of the error. Apparently the vast majority are aware that at the highest levels there are big mistakes.

The questions that I find most interesting are 51, 49, 23. Among the easy questions, 19 is one of the ones that I find most interesting. When I say "interesting" it is because they are more different from other standard problems and require resolution methods that are also different from traditional paths. 19 is not quite like that, as it is simple, but it has some interesting peculiarities for the difficulty level it is at.

Jacobsen: Pragmatically speaking, for really good statistics, what is the ideal number of test takers? You can't say "8,000,000,000".

Melão Jr.: The method I describe in the 2003 Sigma Test standard has a list of important advantages compared to other methods. One of these advantages is enabling more accurate standards based on fewer samples. This happens for a simple reason: in the theoretical normal distribution, rarity decreases rapidly. As measured IQ becomes higher, the addition of a few IQ points implies a large increase in the level of rarity, and test questions are not naturally adjusted to keep pace.

For scores below 140 and especially below 130, IQ scores generally grow almost linearly with the raw scores, and this tracks reasonably well with the theoretical rarity corresponding to each score. But for much higher scores, the gain of 2 or 4 points in the score should not add up to even 1 point in the IQ, because that 1 point in the IQ would imply a very large increase in rarity. In practice, however, IQ scores continue to grow almost linearly with raw scores even for IQs above 140, 150, 180…

The real problem is not in this almost linear growth, but in believing that the real distribution of scores continues to adhere to a normal distribution for scores well above 130, because this obviously does not happen. The number of people with IQs above 200, sd=16 is much higher than would be predicted based on the hypothesis that IQs are normally distributed across the spectrum. When scores are standardized using the method used by Wechsler, the scores are forced to fit a normal distribution, but this only happens within the range determined by the size of the sample used in the standardization (generally 2000 to 3000 people).

Between 70 and 130 the "natural" distribution of scores is very similar to a normal one, and with a "push" it is possible to force scores from 130 to 150 to also be normalized, but in a sample with 3000 people from a non-selected population It is not possible to push scores above 155 close to normal and the distribution collapses. But even if it were possible to use a sample with 8 billion people and push all the scores to the predicted theoretical rarity positions, this would not help at all, it would only expand the distortion by widening the range in which the scores lose intervalarity.

Wechsler's idea of standardizing scores was interesting and would be good for solving some problems, but it creates other problems. In Measurement Theory, whenever possible, it is important that the variable of interest is on a proportion scale. If not, it is recommended to adopt appropriate transformation methods to place the variable on a proportion scale. Height, for example, is naturally on a scale of proportion. IQ measured by the relationship between mental

and chronological age is naturally on a scale similar to a ratio scale. But when Wechsler put his finger on it, he distorted most of the scores to “fix” the problem of IQ variation with age and the wider standard deviation for children.

One of the appropriate solutions for this is the one I propose in the 2003 ST standard, with an updated version in 2022 in this article **https://www.sigmasociety.net/escalasqi** , with a complete reformulation of the standardization method, generating scores on a scale of proportion (antilog of a proportion scale), correcting rarity levels to realistic values and allowing more accurate normalizations with smaller samples, in addition to other advantages.

We can make an analogy with height or chess. First with height: if you try to estimate a person’s height based on rarity level, you will need gigantic samples to measure above 2.10 m and you will still have serious distortions in the results. But if you use a tape measure, a measuring tape, a Leica laser gauge or any other tool for measuring length, you standardize the scale intervals and eliminate the need for large samples.

Chess example: to measure Carlsen’s strength at his peak (2882) with reasonable accuracy and precision based on his results against opponents rated 1000, hundreds of thousands of games between them would be necessary, because the theoretical probability is in favor of Carlsen in a approximate ratio of 50,000:1, so with 100,000 games there would be an expectation of only 2 points for the player with rating 1000. If the player with 1000 scored 1 or 3 points, the error would be large in relation to the 2 points expected, with great uncertainty in measure. It would need a sufficient sample for the player with 1000 to get at least a few dozen points, and for that the sample would require a few million games of him against Carlsen, making it unfeasible.

However, it could introduce players with 1500, 2000 and 2500. The one with 2500 would play 1000 games against the one with 2000 and another 1000 games against Carlsen. The 2000 would play 1000 against the 2500 and 1000 against the 1500. The 1500 would play 1000 against the 2000 and 1000 against the 1000. This way, with a few thousand games it would be possible to achieve a more accurate and precise estimate for Carlsen’s rating, because the expected probabilities in the 500 point intervals are about 94.68% points for the strongest, so there would be a few dozen points for the weakest in each match.

Generalizing the same idea, instead of players with 1000, 1500, 2000, 2500, it could include several players with different ratings playing against each other, using something like the Swiss Pairing System, so that players of similar strengths prioritize clashes with each other, and this would optimize the accuracy and precision of the measurement, without needing a huge number of matches. With players with ratings varying from 100 to 100 points covering the range of 1000 to 2800, and a network with a few hundred matches between them, it would be possible to make a more accurate estimate than if millions of matches were played placing the player at 1000 playing directly against Carlsen.

This is only possible because the method for calculating chess ratings uses the Rasch system, adopted by Arpad Elo. If you tried to evaluate the strength of players based on rarity or percentile, it wouldn’t work and you would need a very different path and with much larger samples.

For this to work with IQ tests, the standardization method needs to be as I described in the 2003 standard, which also uses a Rasch-like model. In this way, the calculation is essentially the equivalent of treating each test item as an opponent in Chess. Solving each item means "winning". The difficulty of the items is equivalent to the strength of the opponents. And for everything to make sense, the approach I give to the problem with the concept of "potential IQ" is necessary.

With this, it can be measured at very high levels with relatively small samples. There is also a more detailed description in the book "Chess, 2022 best players of all time, two new rating systems", in which I discuss several additional details, including the problem of the draw, which in the Chess Elo system is inadequately valued. ed as "0.5", without the necessary adjustments to preserve the consistency of the method.

The problem with the draw value is because the Rasch model used by Elo was created for dichotomous variables, but Chess is trichotomous. Arpad Elo tried some fixes, but couldn't find a good solution and surrendered to simply awarding 0.5 for the draw. There is a 2015 study by Miguel Ballicora that attempts to assign a "fair" value to the draw, and represents an advance compared to the Elo system, but it still incurs several other errors. In my book, I analyze this subject in detail.

Jacobsen: What tests and test builders have you found to be good?

Melão Jr.: I will try to give a generic answer, which complements part of the comments I have already made in the introductory texts for Sigma Test Extended and Sigma Test Light (I also recommend reading these, as a complement). I see 3 main problems (I could divide the problems into 4, 5 or 6 groups, or another number, but in this case I believe that 3 allows an adequate description).

1. Inadequate construct validity, especially at higher levels.
2. "Naive" and inflated norms for scores above 135, with progressively greater distortion in higher scores.
3. Inadequate difficulty.

I could also mention other problems, such as leakage of solutions, retests with fake names, etc. But I will focus on the 3 above.

Good tests that do not fit into one or more of these problems are rare. Furthermore, there are tests with even more serious problems, such as standards based on 1 or 2 people, and even based on 0 people. In some cases, it is very difficult to start standardization with 0 people, but it would be more prudent to estimate a conservative initial norm and eventually correct upwards (after collecting empirical data), however what is most often observed is the opposite.

Therefore, good tests are those that do not incur these problems, that present a sufficiently large number of items with different levels of difficulty in order to measure correctly in each IQ range, preserving construct validity at each level.

Another point to consider is that a test may be suitable for a certain IQ range, but not for a different range. WAIS is a good example. Although it has several flaws, it generates scores that are very close to correct in the range of 85 to 115, and reasonably correct in the range of 75 to 125. It still generates acceptable scores between 70 and 135. Above that, the errors are already worrying. The Power Test can measure well between 110 and 150, and still generates reasonable results up to 160.

Jacobsen: What did you learn from doing this test and its variants?

Melão Jr.: Psychometrics uses some tools that are widely used in other areas, but it also has its own tools, which are rarely used in other areas. I ended up learning some new statistical tools, in addition to developing others.

Jacobsen: Thank you for the opportunity and your time, Melão.

Melão Jr.: I thank you for the reminder and the stimulating questions!

Matthew Scillitani on Divine Psychometry

2024-08-01

Matthew Scillitani, member of the Glia Society and Giga Society, is a software engineer living in Cary, North Carolina. He is of Italian and British lineage, and is fluent in English and Dutch (reading and writing). He holds a B.S. in Computer Science and a B.A. in Psychology. You may contact him via e-mail at mattscil@gmail.com. Scillitani discusses: Divine Psychometry; “a journey to Paradise”; the origin of the title of the test; number of test items; verbal at 20 questions; language a barrier with English as a basis; three easiest test items; three hardest test items ; roadblocks test-takers tend to make; a good numerical test; a good verbal test; a good spatial test; help with this test; sample size of this test; tests and test constructors; and learned from making these tests.

Scott Douglas Jacobsen: You did make a test: Divine Psychometry. What was the general feedback on it?

Matthew Scillitani: Mixed. A few people who tried it told me that it was interesting, enjoyable, or even beautiful. I’ve also had people tell me that it was a little boring or that the problems weren’t as satisfying as those found in Paul Cooijmans’ tests. Since Paul’s the greatest I.Q. test and puzzle constructor of all time, that’s understandable, though.

Jacobsen: It warns, "While this test constitutes a journey to Paradise, it does include an ever so short stay in Purgatory, sorry." Who wrote that?

Scillitani: Paul did. There's a good reason why, but you'd have to see the test yourself to know.

Jacobsen: What is the origin of the title of the test?

Scillitani: I had just finished reading Dante's *Divine Comedy*, and thought it would be fun to make an I.Q. test loosely based around that. Hence, Divine Psychometry.

Jacobsen: The number of test items is 32 with 20 verbal, 8 numerical-symbolic, and 4 visuo-spatial. Why this ratio?

Scillitani: At the time, I enjoyed verbal problems the most, both creating and solving them. So, I heavily focused on verbal items only because it was more interesting for me as the test constructor. Nowadays, I like numerical and spatial problems about the same as verbal.

Jacobsen: Does verbal at 20 questions bias the structure of the cognitive profile tested?

Scillitani: Probably. Someone who does better at solving verbal problems than other types may perform better. Sometimes that doesn't happen, though. For example, Psychometric Qrosswords is a purely verbal-looking test but its numerical g loading is slightly higher than its verbal g loading. This seems counter-intuitive, but it shows that what a test appears to measure isn't necessarily what it's exactly measuring.

Jacobsen: Is language a barrier with English as a basis? If so, how so?

Scillitani: Non-English speakers should still be able to do pretty well on my test. Maybe they'll lose a single raw point at worst, but having the test in English shouldn't act as a barrier. I've taken several verbal tests in foreign languages and my scores on them are always around what I'd score on an English test.

Jacobsen: What do you consider your three easiest test items in it?

Scillitani: To keep the test secure, I can't say. However, what I'd think were the easiest items may not actually be the easiest items. What we intuitively think is easy or hard isn't always so.

Jacobsen: What do you consider your three hardest test items in it?

Scillitani: Same as above. There are definitely some items that I made extremely hard but I can't say which, and as far as I know they ended up being much easier than I expected.

Jacobsen: What are roadblocks test-takers tend to make in terms of thought processes and assumptions around time commitments on these tests?

Scillitani: Most test-takers tend to perform much better than they think they can if only they pace themselves. That is to say, work on the test a little at a time, take breaks, and put the test aside for a while before they submit answers. Cramming a test is going to exhaust the mind just like sprinting exhausts the body. If we try to solve an extremely hard test in a matter of a few hours, it rarely ends well, and we'll probably be left with a headache and a low score.

Aside from pacing, waiting a while before submitting answers is good practice. Almost every Glian I've spoken to has said that they submitted a test only to find more answers right after submitting to the scorer. I've experienced that myself a few times too. It's better to finish a test, wait at least a few days (or even a few weeks or months!), re-examine your answers, and then finally submit. You may find that a few of your answers change for the better after a break and review.

Jacobsen: What goes into making a good numerical test question?

Scillitani: Aside from the use of numbers, I'd say that good numerical test items (1) don't rely on any advanced mathematics knowledge to solve and (2) can't be brute-forced. If any problem, numerical or not, can be solved by guessing over and over or by writing a script, it's a bad problem.

Jacobsen: What goes into making a good verbal test question?

Scillitani: Ensuring the problem requires little to no knowledge to solve. Good verbal analogies and word association problems should follow strict logic, leading the smart candidate to an "aha!" moment, and not need the candidate to spend hours hunting down obscure knowledge.

Jacobsen: What goes into making a good spatial test question?

Scillitani: Good spatial test questions can involve (1) rotation, like showing different sides of an object and having the candidate draw the missing view, (2) showing the candidate images and asking for the common association between them, or (3) having the candidate select or even draw the next shape in a series or analogy.

Jacobsen: Did anyone help with this test? If so, how?

Scillitani: Paul Cooijmans helped a lot by giving critical feedback on every item to improve their quality and formatting the test to reduce file size and make it cleaner overall.

Jacobsen: What is the current sample size of this test? What is the highest score so far?

Scillitani: Fewer than 16 people have taken it, though I don't know the exact number. Probably 12-15 as of this interview. The highest score achieved so far is 32/32 for an I.Q. of 190.

Jacobsen: What tests and test constructors have you considered good?

Scillitani: Paul Cooijmans, Mahir Wu, Ron Hoeflin, and Kevin Langdon are the best test constructors in my opinion. My favorite tests are **Dicing with death**, **Narcissus' Last Stand**, and **The Smell Test** by Paul Cooijmans, as well as **N-World** by Mahir Wu.

Jacobsen: What have you learned from making these tests?

Scillitani: I learned that making good tests takes time, problems shouldn't rely on knowledge to solve (or if they do, the relevant info should be easy to find), and to make a truly good puzzle is more art than science.

Jacobsen: Thank you for the opportunity and your time, Matthew.

Scillitani: Always a pleasure!

Iakovos Koukas on Understanding IQ Test Scores

2024-08-08

Iakovos Koukas is the President and Founder of THIS High IQ Society, 4G High IQ Society, BRAIN High IQ Society, ELITE High IQ Society, 6N High IQ Society, NOUS High IQ Society, 6G High IQ Society, NOUS200 High IQ Society, GIFTED High IQ Network, GENIUS High IQ Network, GENIUS Initiative, GENIUS Journal, IQ GENIUS platform, and Test My IQ platform. He is the author of the GIFT High Range IQ Test series, the GENE High Range IQ Test series, the VAST IQ Test series, and the VICE IQ Test series. He was won the WGD Genius of the Year 2015 Award for Europe, the VEDIQ Guild Intellectual Leader of the Year 2019 Award, and the Global Genius Directory Award of the Year 2021, for his contributions to the global high IQ community.

After taking an intelligence test on our website, you might need an interpretation of your IQ score to understand whether your IQ score is low, average, or high. Since all IQ tests are comparative, your IQ score is always determined in comparison to the scores of other people who took the same test, and most of them earn a score in the average range.

The average IQ score is 100, with a standard deviation of 15, and all IQ scores follow a normal (or Gaussian) distribution, also known as Bell Curve, which is symmetrical around its mean. An IQ score of 100 indicates that the test-takers performance on the given test is at the mean or median level of performance in the statistical sample of test-takers of about the same age used to

norm the test. An IQ score of 115 means an intellectual performance of one standard deviation above the median of 100, a score of 85 performance, one standard deviation below the median, and so on.

Some interesting statistics about IQ scores: 68% of the population scores between 85 and 115, 95% scores between 70 and 130, 99% scores between 55 and 145, and only 2% scores below 70 or above 130. Note that IQ scores may vary depending on how well you rested on the day of your testing and how well you were able to focus and concentrate on the test items during the test without having any distractions.

The IQ classification table below provides a clear overview of all the IQ score ranges, distinguishes different categories based on IQ scores, and helps you better understand your cognitive abilities. An explanation for each IQ Classification is also provided.

IQ Range	IQ Classification
180-200	Profoundly Gifted or Extremely Genius
160-179	Exceptionally Gifted or Highly genius
146-159	Highly Gifted or Genius
130-145	Gifted or Near Genius
120-129	Superior
110-119	Above Average
90-109	Average
70-89	Below Average

IQ Classification: Below Average Intelligence (70 – 89)

Your IQ is below the average, which is 100. Your cognitive abilities allow you to overcome most everyday challenges. You might have experienced some difficulties studying in school, and you aren't probably pursuing an academic career, but you enjoy practical things and occupations. Typical occupations with this level of intelligence are laborers, gardeners, factory workers, and farmhands.

IQ Classification: Average Intelligence (90 – 109)

Your IQ is average, just like the IQ of most people. Your cognitive abilities allow you to overcome almost all everyday challenges. You may succeed in many professional fields, and a specific range of academic studies is possible depending on your motivation. Typical professions

with this level of intelligence are carpenters, shopkeepers, mechanics, electricians, cooks, police officers, truck drivers, and machine operators.

IQ Classification: Above Average Intelligence (110 – 119)

You have a higher IQ than 74% of the population. You have above-average cognitive abilities, which allow you to pursue a successful career in many academic fields and succeed in the profession of your choice. You have a high level of problem-solving and abstract reasoning skills, and you can recognize patterns and details. You enjoy reading books and having interesting intellectual conversations. Typical professions with this level of intelligence are foremen, schoolteachers, nurses, managers, psychologists, and sociologists.

IQ Classification: Superior Intelligence (120 – 129)

You have a higher IQ than 90% of the population, and you possess superior intelligence. You have superior cognitive abilities, which will allow you to pursue an exceptional academic career in a wide range of academic fields and succeed in the profession of your choice. You have a very high level of problem-solving, pattern recognition, and abstract reasoning skills. You have a superior ability to recognize patterns and details, and you enjoy reading books and having interesting intellectual conversations. Typical professions with this level of intelligence are pharmacists, accountants, biologists, chemists, lawyers, physicians, general managers, civil engineers, mechanical engineers, and computer scientists.

IQ Classification: Gifted or Near Genius Intelligence (130 – 145)

You have a higher IQ than 98% of the population. You possess exceptional cognitive abilities which allow you to have a notable professional and academic career in the field of your choice. You have an exceptionally high level of problem-solving, pattern recognition, and abstract reasoning skills. You can easily recognize patterns and details, make connections between abstract concepts, and you enjoy reading and writing books, having interesting intellectual conversations, solving logical brain puzzles, and philosophizing. Typical professions with this level of intelligence are mathematicians, physicists, professors, and researchers.

IQ Classification: Highly Gifted or Genius Intelligence (146-159)

You have a higher IQ than 99.9% of the population. You possess supreme cognitive abilities which allow you to make a notable impact in any professional or academic field. Because of your supreme level of problem-solving, pattern recognition, and abstract reasoning skills, you have the potential to create whole new fields of interest and research. You can very easily recognize patterns where other people cannot and make connections between highly abstract concepts. You enjoy reading books, writing academic papers, having interesting intellectual conversations, solving challenging logical puzzles, and philosophizing. Typical professions with this level of intelligence are university professors and research scientists.

IQ Classification: Exceptionally Gifted or Highly Genius Intelligence (160-179)

You have a higher IQ than 99.99% of the population. You possess exceptionally high cognitive abilities, which allow you to make a significant impact in any professional or academic field. Because of your exceptionally high level of problem-solving, pattern recognition, and abstract

reasoning skills, you have the potential to create whole new fields of interest and research and make significant scientific breakthroughs. You enjoy recognizing highly abstract patterns in anything, reading books, writing academic papers, having interesting intellectual conversations, solving very challenging logical puzzles, theorizing, and philosophizing. Typical professions with this level of intelligence are distinguished university professors and research scientists.

IQ Classification: Profoundly Gifted or Extremely Genius Intelligence (180-200)

You have a higher IQ than 99.99999% of the population. You possess extremely high cognitive abilities, which allow you to make a significant impact in any professional or academic field you choose. Because of your extremely high level of problem-solving, pattern recognition, and abstract reasoning skills, you have the potential to create whole new fields of interest and research, solve the most challenging scientific problems, and make significant scientific breakthroughs. You enjoy recognizing highly abstract patterns in anything, reading books, writing academic papers, having interesting intellectual conversations, solving very challenging logical puzzles, theorizing, and philosophizing. Typical professions with this level of intelligence are distinguished university professors and research scientists.

Iakovos Koukas on Intelligence Types and Theories

2024-08-08

Iakovos Koukas is the President and Founder of THIS High IQ Society, 4G High IQ Society, BRAIN High IQ Society, ELITE High IQ Society, 6N High IQ Society, NOUS High IQ Society, 6G High IQ Society, NOUS200 High IQ Society, GIFTED High IQ Network, GENIUS High IQ Network, GENIUS Initiative, GENIUS Journal, IQ GENIUS platform, and Test My IQ platform. He is the author of the GIFT High Range IQ Test series, the GENE High Range IQ Test series, the VAST IQ Test series, and the VICE IQ Test series. He was won the WGD Genius of the Year 2015 Award for Europe, the VEDIQ Guild Intellectual Leader of the Year 2019 Award, and the

Global Genius Directory Award of the Year 2021, for his contributions to the global high IQ community.

If your scores on IQ tests are not high, there is no need to worry. Intelligence comes in many forms, and some cannot be assessed by modern intelligence tests.

One type of intelligence

Charles Spearman, in 1904 made the first factor analysis of correlations between the tests. Spearman observed that children's performance ratings across unrelated school subjects were positively correlated. He suggested that these correlations were the result of an underlying general mental ability that influenced all kinds of mental tests. Spearman proposed that an individual's mental performance is the result of a single general ability factor, which he called g, and many narrow special ability factors.

The g factor (or general intelligence or general intelligence factor) is a psychometric construct that governs all cognitive tasks and abilities. G factor is a variable that summarizes positive correlations among different cognitive tasks and mental tests. One's performance on one kind of cognitive task tends to be comparable to the same person's performance on other kinds of cognitive tasks. IQ, g factor, general intelligence, general cognitive ability, and simply intelligence are terms used interchangeably to refer to what cognitive tests try to measure.

Two types of intelligence

Charles Spearman developed the two-factor theory of intelligence using factor analysis, which includes the g factor of general intelligence, and the s factor of specific cognitive abilities (verbal, spatial, numerical, and mechanical). Spearman developed a procedure named factor analysis, in which related variables are tested for correlation to each other, and then the correlation of the related items is evaluated to find groups of the variables. He tested how well people performed on different mental tasks, such as distinguishing pitch, perceiving weight and colors, directions, and mathematics. When analyzing the data he collected, he noticed that an individual's performance on one kind of cognitive task tends to be comparable to the same person's performance on other kinds of cognitive tasks. Spearman concluded that there is one g factor that influences all cognitive abilities, but also the s factor of specific intellectual abilities (verbal, spatial, numerical, and mechanical).

Raymond Cattell, in 1963 introduced two types of cognitive abilities in a revision of Spearman's G factor concept of general intelligence: fluid intelligence (Gf) and crystallized intelligence (Gc). Fluid intelligence is the cognitive ability to solve novel problems (like number series and shape classifications) by using abstract reasoning and flexible thinking and depends minimally on prior learning and education. Crystallized intelligence (Gc) is the ability to solve problems (like word analogies and word similarities) by using learned methods and knowledge and depends strongly on prior learning, experience, knowledge, and education. The concepts of Gf and Gc were later further developed by Cattell and his former student John Horn.

Three types of intelligence

Robert Sternberg theorized the Triarchic Theory of Intelligence and challenged the concept of g-factor, and took a more cognitive approach, and it's categorized as a cognitive-contextual theory. The three components are called triarchic components. Sternberg associated the processes of the mind with a series of cognitive components, which he called: meta-components, performance components, and knowledgeacquisition components. Sternberg proposed that the basic information processing components underlying the three parts of his triarchic theory are the same; different contexts and different tasks require different kinds of intelligence.

Sternberg separated his theory into the following three sub-theories: the contextual sub-theory, which says that intelligence is based on how the individual interacts with their environment, the experiential sub-theory, which says that there is a continuous sequence of experience from novel to automation to which human intelligence can be applied; and the componential sub-theory, which outlines the various mechanisms that result in intelligence. Sternberg suggested that intelligence is comprised of three parts: practical intelligence (contextual sub-theory), creative intelligence (experiential sub-theory), and analytical intelligence (componential sub-theory).

Practical intelligence is related to finding solutions that work in your everyday life by applying prior knowledge, experience, and common sense. Analytical intelligence is related to academic problem solving, and it's demonstrated by an ability to analyze, evaluate, judge, compare, and contrast. Creative intelligence is related to imagining a solution to a problem or situation, finding a novel solution to an unexpected problem, or creating a beautiful work of art or well-developed literature.

John Carroll, in 1993 proposed the three-stratum theory, which is a hierarchical model with three layers (strata). The bottom level consists of narrow abilities that are taskspecific (e.g., induction, spelling ability), a few broad factors at the intermediate level, which are fluid intelligence (Gf), crystallized intelligence (Gc), general memory and learning (Gy), broad visual perception (Gv), broad auditory perception (Gu), broad retrieval ability (Gr), broad cognitive speediness (Gs), and processing speed (Gt), and at the top a single factor, the g factor, which accounts for the correlations among all cognitive tasks. The three-stratum theory is an expansion of Spearman's model of general intelligence and Horn and Cattell's model of fluid and crystallized intelligence.

The Cattell–Horn–Carroll theory integrates the Gf-Gc model of fluid and crystallized intelligence with John Carroll's three-stratum intelligence model. Due to similarities with the latter, the two theories were merged to form the CHC model. The broad abilities of the Cattell–Horn–Carroll theory are fluid reasoning (Gf), comprehensionknowledge (Gc), quantitative knowledge (Gq), reading and writing abilities (Grw), short-term memory (Gsm), long-term storage and retrieval (Glr), visual processing (Gv), auditory processing (Ga), processing speed (Gs), decision/reaction time/speed (Gt), General (Domain-Specific) Knowledge (Gkn), Psychomotor abilities (Gp), Psychomotor speed (Gps), Tactile Abilities (Gh), Kinesthetic Abilities (Gk), and 9 Olfactory Abilities (Go). The Cattell-Horn-Carroll (CHC) theory of cognitive abilities is considered by modern psychometricians as the most comprehensive and empirically supported psychometric theory of the structure of cognitive abilities.

Seven types of intelligence

Louis Leon Thurstone challenged the concept of a g-factor and developed a model of intelligence centered on "Primary Mental Abilities." After analyzing data from tests of mental abilities, he identified several primary mental abilities that constitute intelligence, as opposed to one general factor of intelligence. The seven primary mental abilities in Thurstone's model are verbal comprehension, verbal fluency, number facility, spatial visualization, perceptual speed, associative memory, and inductive reasoning.

Verbal comprehension is the cognitive ability to understand the meaning of words, concepts, and ideas. Verbal fluency is the ability to use words quickly and fluency in performing rhyming, solving anagrams, and doing crossword puzzles. Number facility is the ability to use numbers to quickly computer answers to problems. Spatial visualization is the cognitive capacity to visualize and manipulate patterns, objects, and forms in space. Perceptual speed is the mental ability to grasp perceptual details quickly and accurately and to determine similarities and differences between stimuli. Associative memory is the ability to recall information such as lists of words, arithmetic and mathematical formulas, and definitions of concepts. Inductive reasoning is the cognitive ability to produce general rules and principles from the presented information.

Nine types of intelligence

Howard Gardner introduced nine types of intelligence: verbal-linguistic, logicalmathematical, visual-spatial, bodily-kinesthetic, musical-rhythmic, interpersonal, intrapersonal, naturalistic, and existential.

Verbal-linguistic intelligence is the mental ability to analyze information, solve problems using language-based reasoning, use words and combinations effectively in communication, think in words, and use language to express and manipulate complex meanings. It is the individual's fundamental ability to use written and verbal language to achieve their goals.

Logical-mathematical intelligence is the mental ability to calculate, quantify, manipulate numerical symbols, carry out numerical and mathematical operations, solve numerical problems regularly, make decisions based on numerical information, consider propositions, use abstract and symbolic thought, sequential reasoning, inductive and deductive thinking patterns, critical thinking, analyze problems, identify solutions, use abstractions, recognize patterns, detect connections, and conduct scientific research.

Visual-spatial intelligence is the cognitive ability to think in three dimensions, solve spatial problems of navigation, visualize objects from different angles and space, recognize faces or scenes, notice fine details, manipulate mental images, and do graphic and artistic work. It is the individual's ability that helps them identify and manipulate visual and spatial patterns and orient in their environment.

Bodily-kinesthetic intelligence is the cognitive ability to manipulate objects and use a variety of physical skills, involves a sense of timing and a clear sense of the goal of physical activity, as demonstrated by athletes, dancers, surgeons, and craftspeople.

Musical-rhythmic intelligence is the mental capacity to discern pitch, rhythm, timbre, and tone and to recognize, create, reproduce music, as demonstrated by composers, conductors, musicians, vocalists, and sensitive listeners.

Interpersonal intelligence is the cognitive ability to understand and interact effectively with other people and be sensitive to other people's moods, feelings, temperaments, motivations, cooperate as part of a group, and have the capacity to note distinctions among others and entertain multiple perspectives.

Intrapersonal intelligence is the mental capacity to have introspection and selfreflection, understand oneself, one's thoughts and feelings, strengths and weaknesses, and use such knowledge in planning one's life.

Naturalistic intelligence is the cognitive ability to discriminate among living things and other features and objects of the natural world, recognize flora and fauna, make a variety of consequential distinctions in the natural world, and use this ability productively.

Existential intelligence is the mental capacity to answer philosophical questions about human existence, such as the meaning of human life, why we die, and how did we get into this world.

Chris Cole on How to Protect High-Range Tests

2024-08-08

Original publication here.

The suspension of the Mega and Titan tests as admissions vehicles for the Mega Society leaves the Society in a difficult position. The explosion of the Internet since 1995 has made it extremely hard to keep test answers secret. Half of the Mega and Titan test answers are easily available on the Internet today. Even if we were to have a new high-range test in hand right now, it would be compromised within a relatively short period of time, perhaps days. In fact, it's unclear how a high-range test would even be normed without rendering it useless in the process. Is this the end of high-range testing, and potentially the Mega Society?

One possible solution would to be to retain the secrecy of the test in the same way the College Entrance Examination Board does, namely, formulate a very large number of questions and have each specific test consist of a small subset of this larger set. Thus the potential cheater is defeated by the need to memorize thousands of problems. In addition, the test is copyrighted and physically protected.

One problem with this solution is that the College Board has a large market for its tests, and therefore can afford to employ hundreds of test designers to write thousands of sample problems, and additional thousands of test takers to verify and norm the problems. Another problem is that it seems to be a lot harder to write a high-range problem than it is to write a mid-range problem.

One possible solution to the first problem would be to use the power of the Internet for good instead of evil, namely, to publish the test over the Internet and let thousands of interested test takers verify and norm the test for free. While this is a cheap way to get a test normed, it works at cross-purposes to the idea of keeping the test secret.

The second problem, thinking of the problems in the first place, might also be solved via the Internet. Perhaps the test problems themselves could be submitted over the Internet. A system could be set up where people who wanted to take the test would be able to, but they could not receive a "certified" test result until they had submitted some quality problems themselves.

However, experience with this kind of self-generating content over the Internet does not lead to optimism. Quality suffers. Various "political" agendas tend to crop up and mix in with the effort, contaminating the outcome. This has led to several failures, notably Internet dictionaries, encyclopedias, etc.

There is an art to good test design, and the market for high-range tests will support relatively few artists. How can we leverage their efforts?

In looking at many tests, there is a certain pattern that appears. It is possible to classify the problems into groups. For example, Ron Hoeflin has a group of problems about cells formed by intersecting various solids such as spheres, cubes, etc. The solution to one member of this group (say, three cubes) does not help much in the solution of another (say, two cones and a sphere). Yet it might be the case that there is an underlying mathematics that yields the answers to all of the problems in the group. Then a very large number of problems could be generated, where the solution to one problem would not help in the solution of another. This would be ideal for creating an on-line test, because cheating would be impossible.

One difficulty would be in norming such a group of problems. It is usual practice to norm a problem by having a large number of people try exactly the same problem. If the problems were different, how could the test be normed? One problem in a group might be more difficult than another.

The answer to this is twofold: first, it is not true that a given problem has a specified difficulty. The difficulty of a problem is in the eyes of the beholder. What norming does is establish a distribution of difficulties over a sample population, which is an estimate of the distribution of difficulty over the entire population. Thus the real issue is to control the error bars around the estimated difficulty. A problem is rejected if the error bars are too large. Similarly, a group of problems would be rejected if its error bars are too large. The "art" is to select groups that have small error bars.

The second answer to this is to observe that an IQ is not estimated based upon one problem alone; there already is a group of problems involved, namely, the entire test itself. So what we are discussing here is the idea of estimating an IQ based upon a set of problems selected from a large normed set, versus estimating an IQ based upon a set of problems selected from a set of normed groups of problems. Either way, there is an inescapable statistical inference being performed; it's all about propagation of errors.

Another objection to the idea of groups of problems with an underlying mathematical solution is that it might be possible to learn the underlying solution and thus learn how to answer all of the problems in the set. If the underlying mathematics is trivial, this is indeed a weakness. However, it might be that the underlying mathematics is sufficiently complicated that it is easy for a computer to work out, but difficult for a human to work out. Better yet, it might be a one-way or trapdoor function, such as occurs in many cryptographic systems. For example, the Allies during World War II had working copies of the Enigma cipher machine long before the war started, yet they were unable to crack the wartime coded correspondence without cribs, bombes, and a lot of espionage.

As a concrete example, consider problem 30 on the Mega Test. For those without the test at hand, this is the problem where three board positions in some game were given, and you had to figure out the fourth board position. Actually, the first half of the problem was to figure out that the figures shown were board positions in a game that was being played optimally. Even after figuring this out, however, it was a challenge to figure out what the underlying rules of the game were, and to deduce what the fourth position had to be. Now, this one problem could be expanded into a group of problems, by varying the underlying rules of the game and using standard alpha-beta pruned game tree search to find board positions that are unique and lead to simple answers. Even if a test taker know this was what was going on, it would take a similar level of mental effort to deduce the rules from the board positions in each case. And the solution for one set of rules would be of little help in the solution for another set. The size of the board, the number of different pieces, even the movement rules could be varied without greatly affecting the difficulty. A large group of problems with similar difficulty could be created, a group that, according to Grady Towers' item analysis, is one of the best problems on the Mega Test.

Daniel Shea, M.Sc., the Adaptive IQ Test

2024-08-15

Daniel Shea, M.Sc. is the founder and CEO of Chatoyance. Shea possesses a Master's degree in Computer Science from the University of New Hampshire, with several years of industry experience in software engineering. He has published freelance articles on foreign exchange market strategy analysis and has published software analyzing fractals in the foreign exchange markets. Leveraging his experience with software design and financial markets, he started Chatoyance with the intent of transforming the way independent investors approach the foreign exchange market. Shea discusses: interest in test construction; the earlier tests and Chris Cole and Dean Inada; the origin *and* inspiration; Cole and Inada; training in general statistics and software engineering; skills and considerations; help with problem schemas, adaptivity, user interfaces, and renorming; verbal problems and replicability across other problem types; roadblocks test-takers tend to make in terms of thought processes and assumptions around time commitments; the most appropriate means by which to norm and re-norm a test; the Adaptive IQ Test website; tests and test constructors; and the making of a test.

Scott Douglas Jacobsen: When did this interest in test construction truly come forward for you?

Daniel Shea: My involvement came about from conversations with Chris Cole and Dean Inada. There had been an effort to implement an adaptive, generative test many years ago, but it reached a point where conceiving of new high-range questions became increasingly difficult and there were some technical challenges in actually coding a platform to take such a test. Since I had some background on the technical side, I offered to assist.

Jacobsen: What were the general realizations about the earlier tests, e.g., The Mega Test, The Titan Test, The Ultra Test, and The Hoeflin Power Test, of Ronald Hoeflin (Mega Society), and then the need to work in coordination with others for you, i.e., Chris Cole and Dean Inada, to develop a more dynamic test? This form of test development began before you.

Shea: These tests, and other high-range tests available today, are untimed and unsupervised, which introduces many self-evident problems, chief among them being that people will leak answers or collaborate with others. Some of these issues may have been less prevalent at the time these tests were originally constructed in the 1980s and 1990s, but for several years now, many of the answers to these tests have been made available on various message boards or Usenet groups. In some instances, the answers are incorrect or there are multiple answers floating around which muddy the waters, but this is not always the case.

A test should not be entirely discarded just because one or two answers have been leaked. On the other hand, if enough answers have been leaked that one could achieve a sufficiently close score to a given society's cutoff, that society may need to take a vote on whether to continue to allow the test to be used for admission. There is an ongoing effort to identify tests that have been compromised to such a degree, but that judgment call is not an exact science.

Much of the background on the motivation for a dynamic test has been covered in Chris Cole's September 2001 article "How to Protect High-Range Tests" in Noesis #155. To quote, "In looking at many tests, there is a certain pattern that appears. It is possible to classify the problems into groups. For example, Ron Hoeflin has a group of problems about cells formed by intersecting various solids such as spheres, cubes, etc. The solution to one member of this group (say, three cubes) does not help much in the solution of another (say, two cones and a sphere). Yet it might be the case that there is an underlying mathematics that yields the answers to all of the problems in the group. Then a very large number of problems could be generated, where the solution to one problem would not help in the solution of another. This would be ideal for creating an on-line test, because cheating would be impossible." I would probably caution that this does not make cheating outright impossible, but introduces another layer of security.

Jacobsen: Similarly, what was the origin *and* inspiration for joining this small team – the facts and the feelings?

Shea: In a way, the fact that the team was so small made it easier to join. There was a website, mental-testing.com, that had an initial version of the adaptive test, but it was not working at the time that I joined, so the decision was made to rewrite it from the ground up. With greenfield projects in general, there are more degrees of freedom and less rigidity in its development. The ability to make some sort of impact, even if only on a technical level, was appealing. There is also the fact that the Ultra Test and the Power Test, which are the only tests

used for Mega Society admission at this point in time, will eventually be spoiled in their entirety, at which point there will be no viable test for admission without some suitable replacement.

Jacobsen: As an open credit to Cole and Inada, what have been each of their major contributions to the development of the Adaptive IQ Test (2003-present)? (Anyone else, too?) For examples, "How to Protect High-Range Tests" by Chris Cole comments on the difficulties in test questions/high-range tests remaining non-compromised in the internet era, the cost in open-sourcing test creation and norming, and the possibility in designing high-range tests with more foundational principles of math to generate questions (through schemas). Subsequently, "Reply to Chris Cole on Norming High-Range Tests" by Dean Inada commented on something like probability sloping for relative hardness of problems per person and problem. They were discussing, in essence, some foundations for–what would become–the Adaptive IQ Test.

Shea: The background discussed in those articles serves as the foundation for what the Adaptive IQ Test has become in its current iteration. Dean Inada, in his response article, writes "we'll want a better method of norming the tests than simply ranking people by the number of questions they get correctly, since one person may be asked harder questions than another. I suggest a method that tries to estimate for each question the probability of getting it right or wrong as a function of a person's percentile rank in the population, this rank is estimated by multiplying the generally increasing and decreasing functions for the problems gotten right and wrong." The Adaptive IQ Test implements this, modeling an individual curve for the test-taker based on their responses to each administered item and its item curve, and presenting a problem variant accordingly.

Jacobsen: You do not have a formal background in psychometrics. Most people in the high-range construction space do not have a formal background in psychometrics. However, how have training in general statistics and software engineering, i.e., stuff used at Chatoyance, helped with the work on the Adaptive IQ Test?

Shea: As noted, I do not have a formal background in psychometrics. My involvement in the project has been largely technical in nature, drawing on prior general software engineering skills to implement the problem schemas and adaptive component, design the user interfaces for each problem (some may require drawings, some may require filling in a grid, etc.), automate the norming and curving for each item as results come in, and so on. Indeed, the largest challenge has been in conceiving of suitable problem schemas, which I am happy to brainstorm but of course defer to those with a deeper background than my own. Between that and ensuring problem variants are all similarly challenging, progress is ongoing.

Jacobsen: What skills and considerations, in an overview, seem important for both the construction of test questions and making an effective schema for them?

Shea: Among the questions that exist in the current alpha version of the test, these were largely derived from existing problems authored by Ron Hoeflin. The sense was that it was not the problems themselves that were fundamentally at fault here, but rather that it took more effort to vet a sufficient problem than it did for someone to go on to leak it.

With that said, deriving a schema that generates problems of similar difficulty is a challenge, and often requires restricting the degrees of freedom for the generator itself. For instance, the Mega

and Titan item analysis has shown that the interpenetrating solid questions tend to be among the most challenging, but the degree to which they are challenging varies significantly. Consider the three interpenetrating solid questions on Ron Hoeflin's Power Test, which are lifted from the Mega and Titan Tests. There is a notable difference in the difficulty of the interpenetrating cube and tetrahedron compared to the interpenetrating three cubes compared to the interpenetrating two cones and one cylinder. It would not be good practice to include a general schema for any configuration of interpenetrating solids. Rather, you would need to classify these by difficulty and generate them separately. But where does this classification come from? Item analysis gets you started, but at a certain point, you also depend on a sufficient number of people to take the test and get a better idea of the difficulty and signal of each variant.

Jacobsen: How do you help with problem schemas, adaptivity, user interfaces, and renorming? How are the problem schemas developed from the Mega, Titan, and Ultra, tests, e.g., the six sides question from the Ultra Test (problem 45) and grid sequences from the Power Test (problems 32-36)?

Shea: In some ways, it is difficult to discuss particular schemas at length because doing so may reveal the underlying pattern in the process. Many schemas are derived programmatically, while some do not have a proven underlying pattern but are bucketed in the same schema, such as the interpenetrating solid variants discussed prior.

User interfaces are designed according to the requirements of the problem. The most challenging interfaces have been the sixth side problem, which requires drawing on a canvas and scoring the answer in a way that accommodates any orientation of the object, and the three dice problem, whose challenge was less with the user interface per se and more with the backend construction of each variant.

Norming is automatically done after each test has been completed. This also backfills prior test-takers, whose estimates are updated accordingly. In the interest of fairness, there are two metrics presented: the immutable estimate per the norm at the time of the test's completion and the most recent estimate per the latest norm.

Jacobsen: How are verbal problems capable of presenting appropriately challenging problems with variation in type while sustaining similarity of difficulty? Is this replicable across other problem types, e.g., spatial, numerical/quantitative, matrices, etc.?

Shea: Verbal problems in particular have been quite tricky. In the current form of the test, there are trial questions which are presented to the test-taker but do not impact their estimated curve. These trial questions include some, but not all, of the verbal questions. This is in part because verbal problems that have a clean generalization tend to be quite easy to solve. Unlike problems with a more mathematical or logical approach, verbal problems tend to be self-contained, and if generalizable at a high-range, risk producing variants that are far more esoteric than others. This class of problems continues to present the greatest challenge.

Jacobsen: Potentially, what are roadblocks test-takers tend to make in terms of thought processes and assumptions around time commitments on these high-range tests? So, they get artificially low scores.

Shea: In terms of time commitments, at this point, there is no limitation to the length of time that a test may be completed. Historically, it would have been more difficult to enforce, as most high-range tests are made available in their entirety to the public. There are some approaches that are taken to minimize leakage of the questions themselves, such as with Paul Cooijmans requiring test-takers to directly request a copy of the test, though my understanding is that this is done to prevent public discussion of the questions and, in turn, their answers, as opposed to any limitations on time taken to complete the test. Timed tests do allow for a measurement of processing speed to some degree, as well as a standardization of test-taking conditions, but given that these particular tests are already being administered without supervision and in whichever environment the test-taker prefers due to the questions requiring a significant amount of time to answer, timing the test could risk giving an unfair advantage to those who simply have more free time to commit.

As far as thought processes, I do not have enough insight into individual test-takers to make broad generalizations about their personal approaches to these problems. From what I have witnessed myself through discussions with others, there is, perhaps unsurprisingly, a tendency to overthink a question or use complicated reasoning to justify a suspected answer, thereby getting it wrong. Almost every time, the answer is clean; like learning how a magic trick is performed, the question once looked impossible but suddenly seems deceptively simple.

Jacobsen: What are the most appropriate means by which to norm and re-norm a test when, in the high-range environment so far, the sample sizes tend to be low and self-selected, so attracting a limited supply and a tendency in a type of personality?

Shea: Since norms are performed on test completion, the process has little overhead. To accommodate low sample sizes, an initial item curve is provided for questions when known. For example, if a schema is adopted from a prior test such as the Ultra Test, then the item curve for that problem is used as the seed for this test. In some cases, such as novel schemas which do not have a prior item curve from which to draw, the curve starts out flat and is gradually shaped based on the test-taker's answers to other questions.

With these sorts of tests, the low sample size continues to be a problem, but part of this high barrier to entry may be the historical nature of how these tests were administered, between accessibility and cost to score. By making the test available online and without charge, the hope is that this may motivate others to try it out.

As far as the types of personalities that are drawn to high-range tests, I defer to Grady Towers' observations in Noesis #141 regarding the types of personalities that exist across different societies and the corresponding tests used for their admission. Perhaps there is something to be said for stressing both verbal and non-verbal aptitude.

Jacobsen: The Adaptive IQ Test website opens with a series of claims:

This is an online IQ test that contains several innovative features. Here are some reasons to take this test.

1. As you answer more questions, the estimate of your rank in the population becomes more accurate.

2. You see a graph of your estimated rank, not just a single number.
3. You are allowed to skip questions and come back to them.
4. You are automatically asked questions that will help make your estimated rank more accurate.
5. As more people take the test, the graphs become more accurate.
6. There are a number of anti-cheating devices being used.
7. The results of this test may be used for acceptance into various high IQ societies.

Any points of clarification that have been needed on any of these at any time in the past from prospective/actual test-takers or the curious? They can be stated here.

Shea: Some of these points are better characterized as statements of fact about the functionality of the test itself, such as the ability to skip questions. One point to clarify about items 1 and 5 is that the estimate for a completed test may change over time as the test is repeatedly normed. There are plenty of cases across other IQ tests where an individual completes the test and receives an estimate only for subsequent test-takers to receive a lower estimate with the same raw score due to the ceiling being lowered through norms over time, and vice versa. As the adaptive test is normed here, all estimates are updated in unison, preventing this discrepancy between raw scores and percentile estimates across different test-takers. As mentioned earlier, both the estimate at the time of the test's completion and the most up-to-date estimate are presented for completeness.

Jacobsen: What tests and test constructors have you considered good?

Shea: The gold standard for high-range testing has always been Ron Hoeflin's series of tests. These serve as the foundation for much of the existing questions in the current early version of the Adaptive IQ Test. Beyond him, there are many test constructors who have quite novel test items that could be of inspiration.

There is value in multidimensional tests that select for both high-range spatial and verbal problems. I again cite Grady Towers, who wrote of this back in 1998 over the course of several letters published in Noesis #141, where he reflected on the implications for high IQ societies that admit members on the basis of tests that stress both verbal and spatial skills as opposed to one or the other.

Jacobsen: What have you learned from helping in the making of a test?

Shea: It is important to not let "perfect" be the enemy of "good." There will always be shortcomings with any approach. Care needs to be taken to minimize these shortcomings and accommodate them to the extent possible.

Perhaps a second learning is that there is a high-range test vacuum of sorts, and that vacuum is being filled with any number of experimental high-range tests. This is not necessarily an issue in itself, as many of these test items are intriguing and derived from historical best practices, including the very test being discussed here. More to the point, ideally, those with a formal

background in psychometrics would be more involved. I am happy to help where I can, but I also recognize my own limits in this space.

Jacobsen: Thank you for the opportunity and your time, Daniel.

Shea: Thank you for giving me the chance to highlight this project! I feel the need to stress that it is very much in an alpha state and that development is ongoing, but that progress is being made. Special thanks go to Chris Cole and Dean Inada for the decades of work that they put into this long before I arrived, Werner Couwenbergh for his hard work on the interpenetrating solid variants, those who provided input thus far (John Fahy, Nathan Hays, Rick Rosner, and Glen Wooten, among others), and everyone who has provided feedback. I am but a vessel, helping to bring this to fruition where possible.

Bob Williams, The Tools of Intelligence Research

2024-08-15

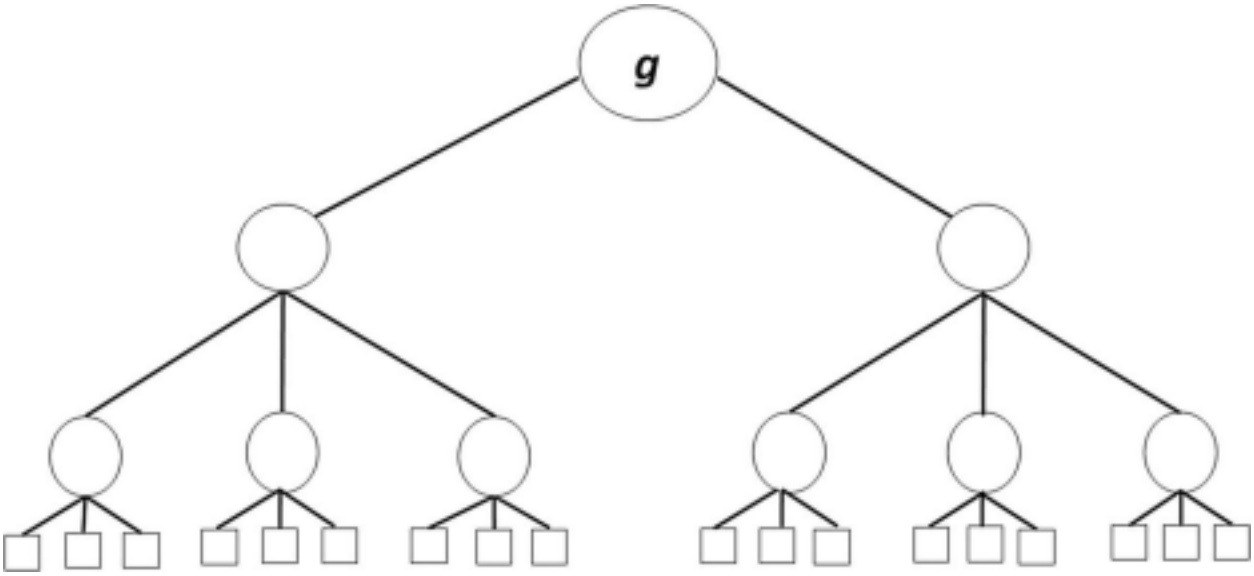
*Original publication **here**.*

Bob Williams is a Member of the Triple Nine Society, Mensa International, and the International Society for Philosophical Enquiry.

The following is a tour through the various methods that have been devised and used to uncover the bits and pieces of insight that make up the present-day scientific understanding of human cognition and its differences among people. The point of this exercise is to identify tools and relationships that are not as well known as the ubiquitous IQ test.

Attempts to understand intelligence go back at least to Sir Francis Galton [1822-1911], who noted the heritability of intelligence, its difference between various populations, and its relation to physically measurable tasks. Following Galton, Charles Spearman contributed new statistical methods, insightful test designs, models of intelligence, and, most importantly, his 1904 discovery of *g* (also referred to as Spearman's *g*, psychometric *g*, the general factor, and *g*). Over the course of the next few decades, *g* languished, while IQ tests were developed, studied, and refined to a point of high reliability and low bias. Numerous well-known researchers contributed models, tests, and understanding that were mostly based on the correlations between test scores and external factors (behavior, physiology, and life outcomes). It was not until Arthur Jensen began to explain the central nature of *g* that intelligence research shifted from earlier models to

converge on g theory. Today, it is difficult to find a research paper that is not about, or constructed from, *g* theory.



Above: Hypothetical example of hierarchical factor analysis

The investigation of intelligence can be sorted into four categories: conventional tests, external measurements with instrumentation, brain imaging, and genetics.

Conventional Tests

Although we are all familiar with some forms of IQ tests, they vary greatly and are designed for a variety of applications. Testing can be done over an age range from toddler to very old. At the young end of this range is the test methodology developed by J. Fagan based on selective attention to novelty (the time toddlers spent looking at new versus familiar faces). His method was predictive of adult IQ (r = 0.59) and adult educational attainment (r = 0.53). The Woodcock-Johnson is one of the broad ability tests that measures a specific number of abilities so that the traditional second-order factors [so-called “group” factors -Ed. Note] of the Cattell-Horn Carroll model will emerge; it claims to measure from age 2 to over 90. The Wechsler, various forms, is also a broad-based test, based on the CHC model, and is considered to be the gold standard (95 percent reliability) by many researchers.

A number of special-purpose IQ-test types have been developed. Some can be given orally to individuals who cannot write (as in an accident victim). Some are designed for speed of administration, taking only a few minutes. These latter group of IQ tests sacrifice range and accuracy for speed and are well suited when a coarse sorting is desired. The Wechsler Abbreviated Scale of lntelligence (WASI) is a well-known example of a test that has been shortened from its full form to achieve this objective. [The WASI is composed of two very highly *g*-loaded subtests (viz., vocabulary and matrix reasoning) as well as the similarities and block design subtests, rendering administration much speedier. A simple vocabulary test may be one of the most effective de facto IQ tests one could give in around ten minutes. Remember that

cultural bias is an empirical question, and cultural bias is orthogonal to cultural load. Cf. *Bias in Mental Testing* -Ed. Note]

As most people have discovered, they are likely to score differently on different tests.This is largely due to uniqueness variance. IQ tests give reasonably close agreement of the latent factor *g* (when it can be computed), but the tests differ in content designed to produce broad ability factors and items that are either specific to the test, or due to random error. Specificity can result from content that is known to the testee (learned material) or is otherwise unique to the test. When a person is trained to take a category of test (teaching to the test), the specificity variance increases, thereby causing the *g* loading of the test to be somewhat lower.

The thing that ties IQ and other ability tests together is known as the positive manifold, which is the strong tendency of a person to score at a similar level on tests of largely unrelated abilities, such as vocabulary and block design. Spearman observed this and created the principle known as the indifference of the indicator, which was intended to point to the universal nature of *g* as a

general ability that appears in all cognitive abilities. Ergo, any test of cognitive ability is predictive of *g*, and all such tests are predictive of the same *g* (meaning that there are not multiple *g* factors). Cognitive ability testing is not limited to IQ tests. There are many tests designed to measure narrow abilities, without an attempt to link the scores to IQ.

Various tests of working memory capacity require the testee to retain representations, while performing tasks that make demands on working memory. He may be given a list of words or letters to remember, separated by a simple task, such as 3 + 5 = 7 (choose yes or no). Then he is asked to recall the list from memory. People are typically able to retain only a small number of representations (4 to 9) in working memory. The simple intermediate math operation effectively flushed out some of the working memory that was used to store the list of memory items. While this category of test is used as a subtest in some IQ tests [Editor's Note: e.g, Working Memory Index on WAIS.], it is also used as a stand-alone tool when working memory is being studied. There are numerous other similar tools that are used for similar purposes.

One of the most interesting special-category tests is the Stroop Color-Word Test. While the test has three parts, it is the third one that demonstrates the Stroop effect. The testee is shown a list of typed color names, but each is printed in a different color ink than the name of the word, (RED is printed with blue ink, etc.). The testee is asked to name, as quickly as possible, only the color of the ink in which each word is printed, while ignoring the name indicated by the printed word.

Red	Yellow	Blue	Green	Black
Pink	Orange	Brown	Gray	Purple
Green	Gray	Black	Blue	Yellow
Gray	Brown	Pink	Orange	Blue
Yellow	Red	Green	Black	Gray
Black	Brown	Purple	Orange	Pink
Purple	Black	Yellow	Red	Green
Orange	Pink	Brown	Gray	Purple

Above: Stroop Color-Word Test

Here is what happens (from Jensen, 2006, *Clocking the Mind: Mental Chronometry and Individual Differences*): "Some individuals are so frustrated by the task requirement that they break down momentarily, while others stammer, stutter, gesticulate, clench their fists, or stamp their feet during this part of the test. Obviously, literate persons are unable to ignore the printed words even when they try their best to do so. Having to suppress their implicit response to the printed word makes it surprisingly difficult to utter the intentional response, viz., the actual color of the print."

The purpose of the test is to measure the executive function or attention (ability to avoid distraction from a task). Research along these lines has linked the executive function, attention, working memory, and *g*. The details of their interdependence are not fully resolved, but they clearly share cognitive resources.

Measurement of Physical Parameters

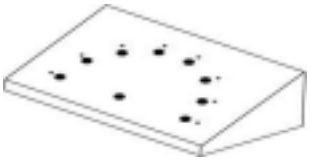
The conventional tests, touched on above, are done with paper and pencil, a computer screen (acting as paper and pencil), or orally. These tests have been used for a majority of the studies of human cognitive abilities. They work and they can be altered to suit the specific mental process that is being studied. Most of them share one significant disadvantage: the tests cannot be scored on a true ratio scale (as is done with most physical measurements, such as force, voltage, mass, etc.). Instead, they have to be scored relative to a selected group of people.

In IQ tests, this is the norming group, and the test is scored by determining the z-score relative to the norming group distribution (IQ = [15 X z score] + 100). The resulting scores are a reasonable approximation of an equal interval scale (as used in the Fahrenheit and Celsius scales).

When physical measurements are used in intelligence research, the results are given on a true ratio scale, such as time, distance, volume, etc. It turns out that a great many of the things that can be measured by instrumentation (including clocks) are linked to IQ test scores and *g*.

Reaction Time (RT)

This measurement is usually done with a Jensen Box and consists of a home button (at the bottom center in the diagram), that the testee holds down, and various target buttons. When the testee sees the stimulus, such as one of the buttons being illuminated, he releases the home button and presses the target button.



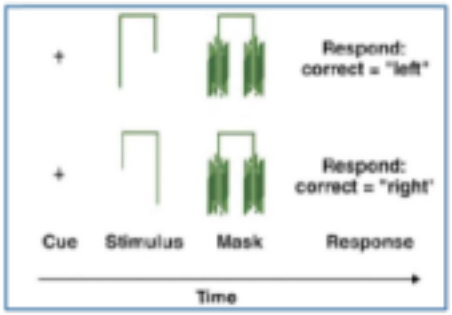
Above: Jensen Box

Reaction time (RT) is measured from the onset of the stimulus to the release of the home button; the time from the release of the home button to the pressing of the target button is the movement time, but is of little value in studying intelligence. Both the RT and the standard deviation of RT are negatively correlated with intelligence, with the latter being somewhat more strongly correlated. RT measurements can be done in connection with a wide range of elementary cognitive tests (ECTs) and can be combined when a battery of these simple tests are given (each requiring less than a second to complete) to produce a measurement of *g* that is approximately equal to the *g* measurement from an IQ test. Each ECT has only a small *g* loading, averaging r = -0.35, but the variances are distinct enough to be added.

Galton performed RT measurements from 1884 to 1893, using a pendulum for the time measurement. His data has been compared to more recent RT studies; it shows that RTs have increased, suggesting a dysgenic effect (explored in detail by M. Woodley).

Inspection Time (IT)

Another widely used chronometric measurement is based on the shortest time that a person can recognize a change in the shape of a projected image. The standard image is somewhat like the letter pi (two vertical lines connected at the top). A cue is given to signal that the test is starting, then the test image is displayed, with one of the vertical lines shortened, then masked. The testee is asked which vertical line of the test image was longer. As the display time is reduced, a point is reached where the testee cannot reliably determine which line was longer. The testee's inspection time is the point where he can achieve an accuracy of 97.5 percent. Again, there is a negative correlation (r = -0.54) between the speed of perceptual discrimination and IQ.



One of the important contributions made by IT was a study by T. Nettelbeck et al. that related to the Flynn Effect. He performed IT measurements for school children from the same school, using the same equipment.

The two sets of data were separated by 20 years. He also administered the same IQ test for the two groups. The expected IQ gain (Flynn Effect) was seen for the test scores, but the IT measurements were essentially identical, thus strongly suggesting that the test score gains were hollow with respect to *g*. I had the opportunity to ask him if there had been any changes in apparent SES, nutrition, or other discernible factors. He said that there was none, and the children were from the same community, school, etc. [Editor's Note: This finding is fascinating and suggests the Flynn Effect could be largely chalked up to practice effects of some kind. Researchers have now found a reversing of the Flynn effect over the last thirty years in various countries, including Sweden, France, and Britain.]

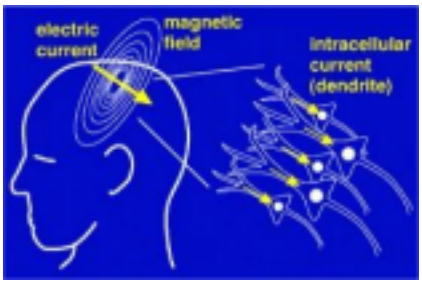
Tachistoscope

IT tests have traditionally been performed by means of a tachistoscope. It has a shutter and can project an image for a precise duration. When computer monitors were first tried for this task, the results were not reliable because of screen characteristics that allowed some people to read screen artifacts. With modern, very fast computer screens this problem has been solved.

Electroencephalography (EEG)

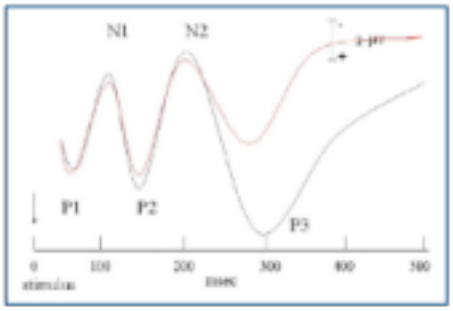
EEG has been widely used for medical diagnostics for head injuries, tumors, infections, and other disorders that relate to the nervous system. The measurements detect electrical activity in the brain by means of electrodes placed on the scalp; these are typically amplified and recorded

on moving paper (creating traces). [Editor's Note: Both EEG and MEG signals are possible because of the electromagnetic laws described by Maxwell's equations, e.g., electrical currents produce an orthogonal magnetic field.] At one time, a good bit of intelligence research was carried out using EEG, but the number of papers reporting it has declined as newer measurement options have appeared.



Depicted above: Ionic current flowing in dendrites, producing an orthogonal magnetic field The magnetic field thus produced is reflected in EEG and MEG readings

A primary focus of interest in EEG has been in the traces made following a specific stimulus. Since the traces contain large amounts of noise, they are repeated many times and averaged to produce the average evoked potential (AEP). The P300 latency, sometimes identified as P3, is one indication of intelligence. It correlates at about r = -0.36 with *g*. Another indication of intelligence is the complexity of the waveform. This is sometimes called string length since it can be measured by laying a piece of string over the wave tracing then measuring its length. Higher IQ is usually indicated by greater string length, but the strongest indication (per T. Bates, et al.) may be the difference in string length between high- and low-attention conditions, which is an indication of neural efficiency.



P300 (P3)

E.W.P. Schafer reported index methods that are based on the amplitudes of the AEP when the stimulus is related to neural adaptability and habituation (see: *The g Factor* for details of the procedures). These methods resulted in correlations as high as r = +0.82 with IQ tests. Although this methodology did not develop a following by other researchers, it demonstrates that *g* is closely related to the electrophysiological activity in the brain.

Other Biological Measures

Intelligence (*g*) is correlated with numerous other biological parameters that can be measured. (Cerebral glucose metabolism is one such measure and will be discussed later.) Nerve conduction velocity (NCV) is inherently related to the speed and efficiency of cognitive activity. NCV has been measured directly in the brain and in peripheral parts of the body. Peripheral measurements (for example, finger to wrist, and wrist to elbow) of NCV correlate with *g* in the range r = +0.41 to +0.46. Although most of these peripheral studies have produced the expected result, some have not, and at least one showed opposite results in men (r +0.63) and women (r = -0.55).

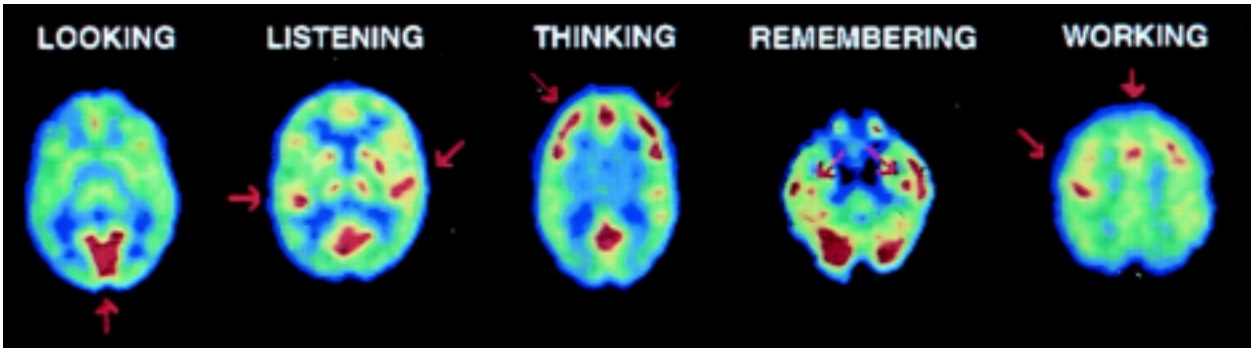
One of the most well-known of these physical measures is brain volume, which correlates positively with intelligence. Before brain imaging technology appeared, brain volume had to be measured by weighing a cadaver brain, or by estimating its volume from the skull volume (taken as the volume of lead shot or mustard seed that it will hold). Another indirect method of measurement is to take the head circumference or multiple measures of length and width to estimate the volume. While head measurements correlate at only r = +0.20 with *g*, the correlation is robust and has been repeated many times with large studies. One of the unexpectedly interesting papers that I have heard presented was Ian Deary's calculation of the IQ of King Robert Bruce (paper presented in Amsterdam in 2007). I think Deary went through the somewhat-complex exercise to teach his students how to deal with data and errors. When it became possible to measure brain volume in a living person, via structural MRI, the correlation coefficient (volume of *g*) of r =+0.40 emerged. This number was later challenged and argued to be lower, but the challenge was subsequently refuted. The best estimate remains close to the initial finding. Brain volume remains an important intelligence parameter, as it relates to intelligence differences between species, between breeding groups (races), and between sexes.

Brain Imaging

Brain imaging technology is to the study of intelligence as the Hubble telescope has been to cosmology. Imaging has appeared in several stages, and each has opened new paths of study and huge gains in the understanding of intelligence.

Positron Emission Tomography (PET)

PET can be used to create images of the brain and various other organs. The thing that is seen as an image is the accumulation of a radioactive tracer (oxygen-15, fluorine- 18, carbon-11, or nitrogen-13) as the tracer is concentrated by the action of the organ being studied. As the tracer decays, it emits a positron, which collides with a nearby electron and causes the emission of two photons. The photons are detected externally.



Above: Positron Emission Tomography with presumed brain states

In the case of brain imaging, the image is effectively an integral of glucose uptake rate. The tracer used is fluorodeoxyglucose, which gives a time resolution of about 32 minutes. Thus, the image produced when a person is asked to perform a cognitive task is an integral over a time span of 32 minutes. The first use of PET to study intelligence was done by Richard Haier (presently editor of the journal *Intelligence*) in 1987. At that time, the cost of a single scan was $2,500. Haier financed the initial work by agreeing to do medical scans in trade for some research scans. His first subjects were given the RAPM (Raven's Advanced Progressive Matrices) during the exam. Raw test scores ranged from 11 to 33 (out of a possible 36). The PET scans revealed the opposite of the expected result. The brighter subjects showed less brain activity (lower glucose uptake rates) than did the duller subjects. This was the first indication that one difference between brains of different intelligence levels was efficiency. The smarter brains solved the problems more efficiently. Decades later, we have numerous other imaging studies, using other technologies that have made similar findings and have added more detail to the initial study. One somewhat-easy-to-find refinement was that all brains show increased activity (effort) as problem difficulty increases, but less-intelligent brains reach a saturation point beyond which they cannot apply additional effort.

Haier also looked at the effect of learning, using the game Tetris. [Editor's Note: Mega Society qualifier and mathematician Solomon W. Golomb's game of pentomino directly inspired Tetris.] Several subjects were given practice sessions with the game (new at that time). They had not seen the game before and were restricted to uniform practice sessions. They improved their play score by a factor of 7. PET scans before and after the learning sessions showed significant reductions in brain activity in some parts of the brain. Haier wrote: "We concluded that with practice and improved performance, subjects learn what areas of the brain not to use, and this results in GMR (glucose metabolic rate) decreases."

PET studies showed the value of being able to measure actual brain activity while subjects were performing mental tasks. The technology was expensive and had the slow 32-minute temporal resolution, so it was displaced when faster, MRI-based machines arrived.

Magnetic Resonance Imaging (MRI)

The first MRI was performed on a human in 1977. The machines are based on the use of a very strong magnetic field (5,000 to 20,000 gauss; the earth's magnetic field measures 0.5 gauss) that is achieved by means of a superconducting magnet. A few years ago, R. Haier told me that there was an MRI machine that used a magnetic field that was significantly higher (ten times, as I recall) than other machines. He said some people complained of headaches and that the brain was warmed – probably causing the headaches. (A recent literature search shows that possibly even more powerful, new MRI scanners have been built. The reason for increasing the magnetic field strength is that it enables the voxel size to be reduced from 1 mm to 0.1 mm.)

MRI

MRI works by imposing an intense magnetic field around the area to be imaged using superconducting magnets. Hydrogen nuclei (protons) spin and have a natural magnetic polarity. When on, the magnetic field causes hydrogen nuclei to snap into axial alignment with the field.

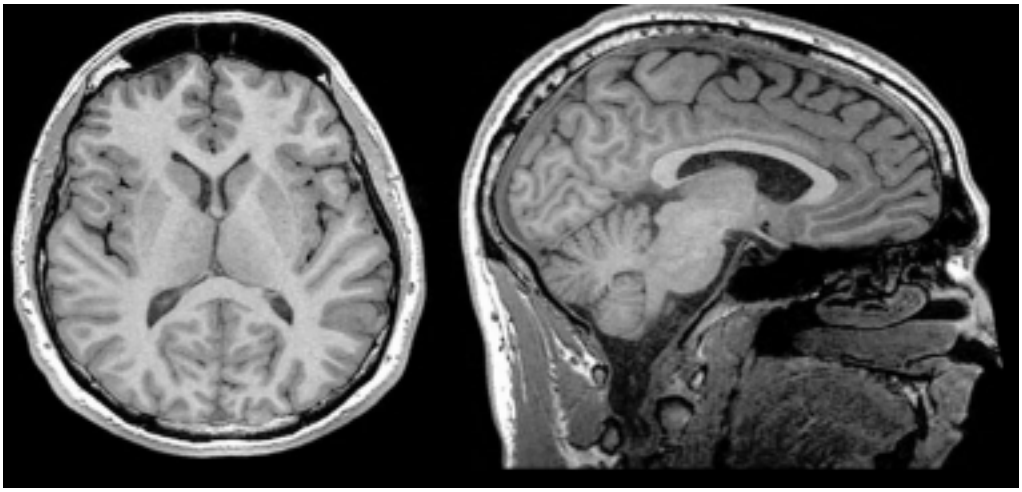
A radio frequency wave is added and is pulsed on and off, causing the nuclei to snap out of alignment and then back in. This shifting of nuclei alignment causes a weak energy release (also a radio frequency wave), which can be detected by the MRI machine (via receiver coils that act as aerials) and used to create an MR image.

Structural MRI (sMRI)

This basic technology (the same as many have experienced in a medical setting) can be varied to allow various specialized forms of imaging. The most basic application for intelligence research is structural MRI, or sMRI. This is essentially a snapshot of the brain, but the image is 3D. It can be rotated and viewed from any angle and can produce a "slice" image of the brain at any depth. Since the image is in 3D, the points are also 3D, unlike the 2D pixels of a digital photograph. The 3D representations are known as voxels.

One of the problems encountered in understanding a brain image is that brains are not identical in size and shape. Yes, they are all generally the same in appearance, just as our faces are similar yet different enough that we can recognize them. A researcher must be able to compare brains, despite their differences. This can be accomplished by a computer using a process known as voxel-based morphometry. The process morphs the MRI data to fit a standard form and smooths the results so that they can be analyzed. For example, an area of great interest is cortical thickness. In order to study it and to compare different brains, the cortex representation has to be smoothed so that the folds are removed and the resulting artificial image retains the dimensions that are of interest, while losing the irregularities that would otherwise make it unmanageable.

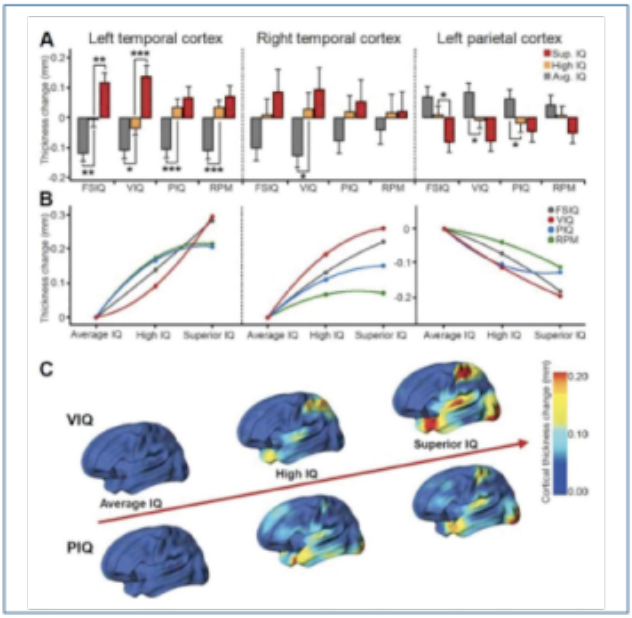
Above: Left image (axial view) and right image (sagittal view) of structural MRI



The cortex contains cortical columns that are vertical structures of variable length and composition. The number of these columns is related to cortical surface area, while their length is a function of cortical thickness. Their relation to intelligence is known primarily by the correlations found in average and local measurements of cortical thickness and in cortical surface area. A good bit of study of cortical thickness (CT) has been related to the NIH (National Institute of Health), e.g., the *Study of Normal Brain Development.*

One finding is that cortical thickness increases in early childhood, then begins a slow decrease around ages 7 to 10 years. When plotted against time, the trajectories of bright children (from longitudinal NIH data) show greater thickness at every age than for less bright children. During the first phase, thickness increases more rapidly in bright children, but exhibits a similar rate of thinning following the peak. This has obviously important significance in the verification of the high heritability of intelligence; the trajectories are set from early childhood. The strongest correlations between CT and IQ are found in the age range of 8 to 12 years.

The figure (below) of CT for different intelligence groups shows that there are differences and that they vary as a function of age. The illustrations of CT as a function of intelligence at the bottom of the figure also show how a brain appears after computer smoothing.



Above: Intellectual domain effects on cortical thickness changes as a function of IQ level. A, Cortical thickness differences between adjoining levels of IQ as affected by intelligence criteria and brain lobes. The superior, high, and average IQ groups were evenly divided according to four intelligence criteria, FSIQ, VIQ, PIQ, and RPM scores. The cortical thickness of each lobe is represented by the averaged value of all ROIs within the lobe. Sup., Superior; Avg., average. *p < 0.05; **p < 0.01; ***p < 0.001, two-tailed t test. B, C, Cortical thickness deviations from the thickness of the average IQ group used as zero reference. VIQ groups are better described by a linear or quadratic function, whereas PIQ groups are better described by a logarithmic one. The brain maps show absolute thickness changes at each cortical point, based on VIQ and PIQ levels.

When the thicknesses of specific locations are correlated against IQ, the results are different for men and women (a surprise to Haier and his team). The highest correlations (gray matter regions) in men were found in posterior regions, especially those related to visual-spatial processing. In women, the IQ-to-thickness correlation was almost entirely limited to the frontal lobes, especially in the language area (Broca's Area). Findings that show sex differences have been frequent, and each strongly suggests the need to keep male and female data separate. Haier made this point to the International Society for Intelligence Research (ISIR) conference in 2006.

Functional MRI (fMRI)

MRI can be used to create images based on molecules containing iron, which is highly sensitive to the intense magnetic fields of MRI machines. Hemoglobin in red blood cells contains iron and iron molecules, thus connecting the fMRI images to blood flow in the brain. When a brain region is cognitively active, it will have greater blood flow, and this will be seen by the fMRI scan. The

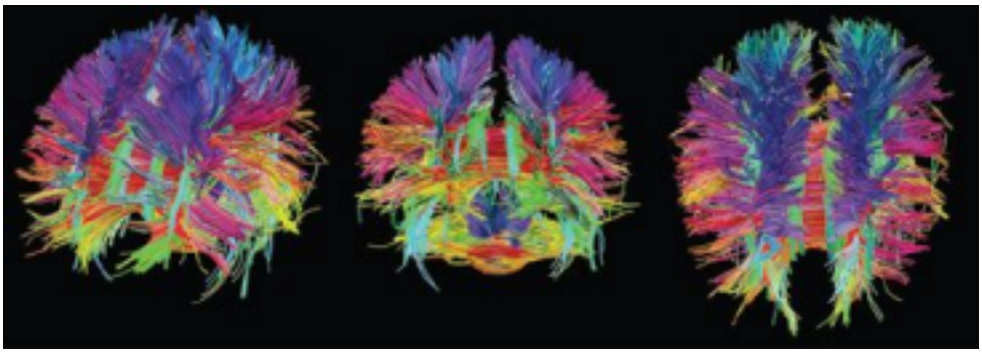
fMRI process is fast, with thousands of images per second and a net resolution that is a span of about 1 second.

One of the applications for fMRI is the study of functional connectivity. When static measurements are made, the information conveyed relates to the function of a given brain region (functional segregation). But as imaging research progressed, brain regions were found to work together, such that a single region is necessarily involved in multiple functions. With fMRI, it is possible to see the connected activities of brain regions.

Using fMRI, it is possible to observe the brain performing a task over a period of time. Various regions show activity (increased blood flow) sequentially, as the brain deals with the task. In a conversation with R. Haier, he mentioned to me that fMRI data were proving to be difficult to use because of the large differences seen between individuals. This is not a problem with static imaging techniques, such as fMRI and diffusion tensor imaging.

Diffusion Tensor Imaging (DTI)

DTI is a different form of structural MRI. It is optimized to image the water content of white matter. The first study did not happen until 2005. Prior to then, white matter was relatively difficult to study. It was possible to measure white matter volumes and to do correlations with that and intelligence (revealing a large sex difference), but the details of how white matter tracts were organized were hidden. DTI has opened a new field of research-brain connectivity (wiring).Among the things that have been found are that the tracts form bands (in some places) that are composed of large numbers of parallel tracts; that each person has tract patterns that are as unique as fingerprints; that the primary cognitive centers are connected by massive highways of tracts, running from the frontal lobes to the parietal lobes; that connectivity is an indicator of IQ.



Above: Diffusion Tensor Imaging

When water movement is detected by the MRI process, it can be quantified as to the degree to which the molecules move in the same direction. This parameter is known as fractional anisotropy (FA) and is higher when the movement vectors are directionally similar. If FA is low, it indicates that the water movement is more diffuse, and this is taken to be an indication of low tissue integrity. Higher FA is a positive correlate of intelligence for both white and gray matter.

Magnetoencephalography (MEG)

Breakthroughs in instrumentation have continued to appear, offering new capabilities. Magnetoencephalography (MEG) is in some regards similar to EEG, in that sensors are placed on or very near the scalp. These highly sensitive superconducting sensors detect magnetic fields associated with neuron activity. The instruments are functional, in the sense of fMRI, but faster; they have a temporal resolution in the millisecond range. The precision of spatial location is excellent – sources can be localized with millimeter precision.

Unlike other methods of brain imaging, MEG is completely passive and is a direct observation of the brain, while other techniques are measuring secondary phenomena (isotope decay, water movement, etc.). MEG is thus totally safe and noninvasive.

MEG

When copiled into a movie, brain activity can be seen as a function of time. This was demonstrated (by Thoma) at the 2005 ISIR conference, showing the brain reacting to a simple optical stimulus. The activation areas appeared to bounce and flow from the extremes of the brain, in much the same way as water waves bounce and reflect when they are confined. When I saw this, there was an immediate revelation as to why something as simple as a light turning on would stimulate activity throughout the brain; this simple event, when measured by RT is significantly correlated with *g*. The video showed that the mental activity was complex and involved most of the brain volume.

MEG remains as a new tool with a limited history for intelligence researchers. It has great promise and is being evaluated by researchers. An example of an MEG movie, made while the subject is solving a test item from the paper-folding task, can be found here: **http://www.cambridge.org/us/academic/** subjects/psychology/cognition/ neuroscience intelligence (select: student resources, then animations, then animation_4.3.mp4).

Genetics

Although Galton observed that intelligence was a family trait, the role of genetics in determining intelligence was not understood for many decades. In the 1960s, even scientists believed that intelligence was largely a product of the environment (books in the home, encouragement to excel in academics, etc.). When Arthur Jensen entered the field, that is exactly what he expected to find, but when he looked at real data, he saw a different story. The result was his 80-page landmark paper: “How Much Can We Boost IQ and Scholastic Achievement?” by Arthur R. Jensen, University of California, Berkeley, Harvard Educational Review, Vol. 39, No. 1, Winter 1969, pages 1-123.

From that point on, Jensen published a huge number of papers and books that addressed the issues related to demonstrating that intelligence is primarily the product of genes, with little environmental variance. Of the environmental variance that is found, it can be divided into the shared and the nonshared environmental factors. The former is that part of the environment that makes us more similar (family), and the latter is that part that makes us more different. There is a shared environmental variance in early childhood, but it vanishes by about age 12, leaving only the experiences people have as individuals (the following factors lower intelligence), such as: injury, disease, exposure to toxins, etc. From early childhood on, the heritability of intelligence increases (the Wilson Effect) into adulthood. By adulthood, the heritability of IQ is 85% and the heritability of *g* is 91%.

Although repeated studies have shown this high heritability of intelligence, attempts to find a single intelligence gene (or a few genes) have failed, despite methodologies that would have found it without doubt. This research has been led by Robert Plomin, who has authored numerous papers on the topic of the genetics of intelligence.

What is going on? The simple answer is that intelligence genes have been found, and each has accounted for only a percent or less of the total variance. As has been the case for other traits, intelligence is the product of hundreds or thousands of variants. For example, height has been shown to be determined by more than 900 variants. The two concepts that relate to this are pleiotropy (one gene affecting multiple traits) and polygenicity (many genes affecting one trait).

Genetic research will hopefully tell an increasingly complete story of which genes are involved, and how. To date, there is an impressive research category known as genome-wide association studies (GWAS). These studies include some with N of much more than 100,000 and at least one that is approximately 1,000,000. The GWAS studies have included genetic clusters that relate to intelligence, educational attainment, and behaviors throughout life. Because of the large N's, the findings are robust, but they show small effect sizes.

A 2017 preprint (**http://www.biorxiv.org/** content/early/2017/07/07/160291) showed 107 independent loci associated with intelligence, implicating 233 genes, using both SNP-based and gene-based GWAS. Further studies will surely appear, and the findings will presumably, if slowly, paint a picture of how intelligence is determined at the molecular level.

Further Reading

For those who are interested in reading original intelligence research papers, there is only one print journal dedicated to this subject: *Intelligence*. It is the official journal of ISIR and is the source of some of the best research papers. Another source that frequently contains top-quality work is *Personality and Individual Differences*. In the area of brain imaging, there are worthwhile papers in *Neuroimage*, *Neuroscience*, and *Cortex*.

The best book and DVD material that is relatively recent:

Haier, Richard J., (2017), *The Neuroscience of Intelligence*, New York: Cambridge University Press. This book is recent and was skillfully written to be easily readable, yet complete with respect to present-day understandings.

Haier, R.J., (2013), *The Intelligent Brain*, The Great Courses, Chantilly, Virginia (3 DVDs).

The first DVD is a review of non-imaging research. It then gets into the very interesting work that Haier and his colleagues have done.

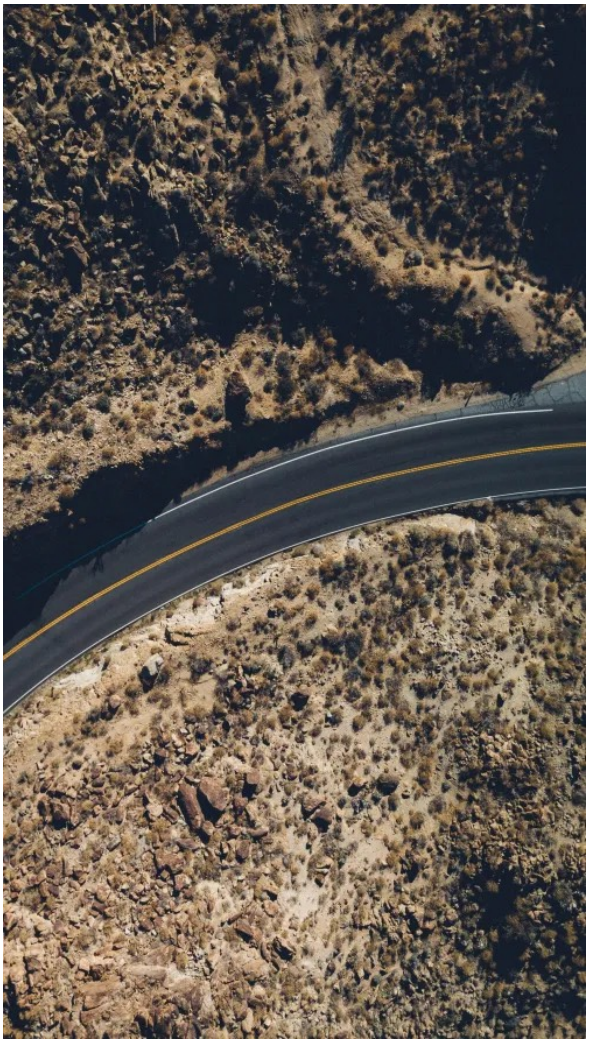
Jensen, A. R., (1998), The g Factor: The Science of Mental Ability, Westport, CT: Praeger.

Written by the most outstanding intelligence researcher of the second half of the 20th century, this book was, and presumably still is, the all-time most cited book in this field.

For those who want excellent and accurate information that is written for public consumption (some exceptions), I strongly recommend the articles and papers by Linda Gottfredson. She has generously made virtually everything she has written available on her web page: **http://www1.udel.edu/** educ/gottfredson/reprints.

Bob Williams, Overview of the Flynn Effect

2024-08-15



Original publication here.

Bob Williams is a Member of the Triple Nine Society, Mensa International, and the International Society for Philosophical Enquiry.

Following WW2, various researchers found and reported secular gains in IQ, but it was not until additional reports appeared in the 1980s that researchers began to look for the cause or causes. It was quickly apparent that the gains were not limited to any group or nation, but the manifestation of the gains was different depending on time and place. For every discovery, there was a different or opposite result in a different data set. Gains have been large, small, variable, and even negative. Some researchers have found that the gains were on *g*, while more have found no *g* loading. Abstract test formats, such as the Raven [Matrices -Ed. Note] have often shown the greatest gains, but gains have also appeared in tests of crystallized intelligence. Some data has shown greater gains for the lower half of the intelligence distribution, while others have shown

greater gains in the top half, and others have shown equal gains at all levels. Hypotheses for the causes have included environmental factors, genetic effects, reduced fertility, and methodological dependence. Two models are discussed.

1. Introduction

The secular rise in IQ scores appeared unexpectedly and has defied explanation. Smith (1942) recorded a gain (in Honolulu) over a 14 year span. Later, Tuddenham (1948) found an increased intelligence when he compared inductee scores for the U.S. Army from World War I and World War II and proposed that the gains might be due to increased familiarity with tests; public health and nutrition; and education [the gains from 1932 to 1943 were 4.4 points per decade.]. He cited a high correlation (about .75) between years of education and the Army Alpha and Wells Alpha tests that he was studying.

The secular gain remained relatively dormant until it was rediscovered by Lynn (1982) while working on a comparison of Japanese and U.S. data. It was then rediscovered again, using American data, by Flynn (1984a,b). The raw score gains did not have a name until Herrnstein & Murray (1994) coined the term Flynn effect in their book *The Bell Curve* (p. 307). ["We call it 'the Flynn effect' because of psychologist James Flynn's pivotal role in focusing attention on it, but the phenomenon itself was identified in the 1930s when testers began to notice that IQ scores arose with every successive year after a test was first standardized." -Ed Note] Some researchers choose to refer to the secular gain as the Lynn–Flynn effect, or use an uppercase FL (FLynn effect) for the obvious reason that they feel Lynn has been somewhat slighted by not including his name.

[FE is the shorthand used throughout the remainder of this overview. -Ed. Note]

Since the early '80s, researchers have found the FE in virtually every group they have examined (Flynn, 1987 and others). They have published a huge number of papers (well over 100) on the gains and possible causes, but the results have been contradictory.

2. Gains

FE gains vary from country to country and over different time intervals, but the gains are usually a fraction of a point per year. As a matter of convenience, the gains are usually given as the number of points gained over a decade and written "ΔIQ." A few typical national gains:

- U.S. ΔIQ = 3 (14 points over 46 years, 1932–1978)
- Estonia ΔIQ = 1.65 (12 points over 72 years, 1933/1936 to 2006)
- Japan ΔIQ = 7.7 (19 points over 25 years, 1940 to 1965)
- Argentina ΔIQ = 6.91 (21.35 points over 34 years, 1964 to 1998).

[Numerous other rates are given in Flynn and Rossi-Casé (2012).].

South Koreans born between 1970 and 1990 gained at about the same rate as did the Japanese (te Nijenhuis, Cho, Murphy, & Lee, 2012). Chinese gained 4.53 points over 22 years (ΔIQ = 2.1) on the Chinese WPPSI (Liu, Yang, Li, Chen, & Lynn, 2012). [WPPSI = Wechsler Preschool & Primary Scale of Intelligence. -Ed. Note] FE gains have been found in both industrialized and

third world nations. The number of countries showing a FE is subject to change, since additions are frequently reported. Kanaya, Ceci, and Scullin (2005) reported 20 nations; Flynn and Rossi-Casé (2012) reported 31.

Teasdale and Owen (1989) examined two samples of Danish draftees, consisting of 32,862 and 6,757 males. They found that the gains were concentrated mostly among the lower IQ levels and concluded that changes in the educational system were driving the score gains. They also performed an interesting test, using Monte Carlo simulations to demonstrate that the FE gain was not caused by a ceiling effect. Flynn and Rossi-Casé (2012) noted that some data sets (they were examining Raven scores) have attenuated SDs [standard deviations -Ed. Note] because of ceiling effects.

Other researchers, including Lynn and Hampson (1986) and Colom, Lluis-Font, and Andres-Pueyo (2005), have found FE gains that were mainly concentrated in the lower IQ levels. This pattern suggests that the gains are related to improving environmental conditions in non-industrialized countries, rural areas, and low-income sectors.

Although it has now been 14 years since Jensen (1998) published *The g Factor*, his discussion of the FE remains current with respect to the items he considered. He reported U.S. gains:

- Raven ΔIQ = 5.69
- Wechsler ΔIQ = 5.2

Performance ΔIQ = 7.8

Verbal ΔIQ = 4.2

These show greater gains on the most abstract tests and subtests, although it is surprising to see the Wechsler as close to the Raven as the above numbers indicate — both being above the usually cited U.S. rate (ΔIQ = 3).

When Jensen examined subtests more closely, he found that non-scholastic test items showed increases at the same time (same test data sets) that scholastic items were decreasing. He noted that this is not what one would expect to see, but this is indeed what other researchers have reported. Jensen examined the SAT for the period 1952–1990 and found the well-known decline. The usual explanation for the decline is that each year more students took the test and most of the additions to the pool of test takers were added below (lower intelligence) the prior group, leading to a decline at the mean. But Jensen corrected for the changes in demographics and showed that 3/4 of the decline was due to the addition of more lower IQ testees, while the remaining 1/4 was a real decline in scores. The ΔIQ loss for the SAT was −5 for the time period in question, while the FE gain was +3. This strongly suggests that the IQ test scores were not reflecting real world gains in intelligence.

2.1. Estonia

Thanks to the work done by Olev and Aasa Must, there is a good bit of information about the FE as it has appeared in Estonia. The messages from their studies are that the FE gains follow

different trajectories in different countries and the factors most likely to be driving those changes are also different.

In the Estonian studies, subtests that needed computation skills and mathematical thinking were unchanged over 60 years. The information subtest declined; verbal subtests showed moderate gains; but there were impressive gains in symbol–number and comparison subtests (Must, Must, & Raudik, 2003).

Must, te Nijenhuis, Must, and van Vianen (2009) examined data over a 72-year span and found a relatively small ΔIQ of 1.65. But when the eight [nine? -Ed. Note] years from 1998 to 2006 were examined separately, the ΔIQ almost doubled to 3 points. The *g* factor loadings were different at the subtest level for each of the three birth cohort groups examined, with the greatest difference between the oldest cohorts compared to the other two relatively recent cohorts.

In recent years, large gains were observed in arithmetic, information, and vocabulary. These gains are opposite from score changes seen in the U.S. and Britain. The authors identified several possible causes: greatly improved education, better nutrition, better health care, and changes in demographics (smaller families).

In 2012, the Estonian data was re-examined at the item level (see Section 4.2.1). The results of that effort are important to the understanding of at least one cause and of an otherwise perplexing difference between Classical Test Theory and Item Response Theory results (see Section 4.9.2).

2.2. South Africa

ΔIQ = 3.63 Whites (same group took two different test batteries)

ΔIQ = 1.57 Indians (same group took two different test batteries)

The FE score gain is stronger for the Afrikaans speakers than for the English speakers (te Nijenhuis, Murphy, & van Eeden, 2011).

2.3. Gains seen in young children

British children aged 6 and 18 months displayed large developmental gains over the period from 1949 to 1985. When measured on the Griffiths Test, developmental quotients (DQ) gained 2.45 points per decade. Similar studies, using the Bayley Mental Scales (Bayley, 1993) were done by other researchers in the U. S. and Australia and show gains of 2.9 DQ points per decade (Black,

Hess, & Berenson-Howard, 2000; Campbell, Siegel, Parr, & Ramey, 1986; Lynn, 2009a; Tasbihsazan, Nettlebeck, & Kirby, 1997). Similarly, Kanaya et al. (2005) reported that elementary school children show FE gains on the WISC that are similar to adult gains on the WAIS. [WISC = Wechsler Intelligence Scale for Children, WAIS = Wechsler Adult Intelligence Scale -Ed. Note] These DQ and IQ gains show a FE that is as large in infants and preschool children as in adults, making education an unlikely explanation for the cause (at least in the data sets examined).

As is already apparent, FE findings in one place do not generalize globally. Cotton et al. (2005) found no FE effect, using the Raven's Colored Progressive Matrices, for a group of Australian

children ages 6–11 from 1975 to 2003; but Nettelbeck and Wilson (2004) found 5 point gain for a range of Australian elementary-grade children from 1981 to 2001.

2.4. Gains in the Raven's Progressive Matrices

The Raven tests have been cited frequently in the FE literature because most samples show particularly large gains on these tests. The Raven and similar tests have shown gains of 18–20 IQ points per generation in many industrialized countries (Flynn, 1999). Dutch gains were 21 points over 30 years (ΔIQ = 7), while urban Chinese gained 22 points between 1936 and 1986, ΔIQ = 4.4 (Neisser, 1998).

Hiscock (2007) found a higher rate of FE gains for the Raven's Progressive Matrices than for the Wechsler and Stanford–Binet tests. He also showed that British Raven scores for birth years from 1877 to 1967 increased steadily, but rolled off over that time span to a possibly flat (no effect) rate for the last 10 year interval.

[The popular Raven's matrices tests – e.g., Standard Progressive Matrices, Colored Progressive Matrices, and Advanced Progressive Matrices – are non-verbal, multiple choice tests which purport to gauge abstract reasoning, i.e., pattern recognition. -Editor's Note]

2.5. Low-end versus high-end gains

As previously mentioned, Teasdale and Owen found that FE gains for Danish draftees were concentrated in the lower end of the intelligence spectrum, suggesting a cause or causes such as improved nutrition, better health care, or increased education. Colom, Andre's Pueyo, and Juan-Espinosa (1998) noted that FE gains were much greater on the Raven's Standard Progressive Matrices (19.2 points over 28 years, ΔIQ = 6.9) than on the Advanced Progressive Matrices (6.75 points over 28 years, ΔIQ = 2.4). They concluded that the cause of the increases probably had a greater impact in the low and medium segments of the intelligence distribution. In a later study, Colom et al. (2005) also found that gains were more pronounced in the lower range.

Lynn and Hampson (1986) reported a low-end gain that was about double the high-end gain, for a British group over the period 1932 to 1982. Similarly, Kagitcibasi and Biricik (2011) found greater gains in Turkey at the low end, over the period from 1977 to 2010. The differences were particularly large (23 points, ΔIQ = 7) for remote villages. Within urban locations, the lower SES groups also showed more gains (7.4 points, ΔIQ = 2.2) than higher SES groups, but these were less than in the remote villages.

The FE is so specific that for every finding, there seems to be an opposite finding. Flynn (1996, 2009) claimed IQ gains at "every level," based on his observation that "score variance remains unchanged over time." His "every level" projection held in a study conducted in La Plata, Argentina, where ΔIQ = 6.3 and showed no bias towards high or low IQ ranges. Flynn extended this observation as meaning that nutrition is an unlikely explanation, since it would presumably apply more readily to gains seen at the lower end, and not throughout the intelligence spectrum (Flynn & Rossi-Casé, 2012). Flynn (2009), cited Sundet, Barlaug, and Torjussen (2004)) as an example in which IQ gains were concentrated in the lower half of the IQ spectrum, while height gains were mostly in the upper half, pointing out that this combination is inconsistent with the nutrition argument.

Colom, Flores-Mendoza, Francisco, and Abad (2007) examined data for Brazilian children covering a span of 72 years. They found that the FE gains were greater for urban samples than for rural samples and concluded: “Whatever the causes of the increase, they act more intensively for more intelligent children.”

Ang, Rodgers, and Wänström (2010) computed FE gains from the National Longitudinal Survey of Youth (NLSY) data, which include scores from the Peabody Individual Achievement Test (PIAT); the math portion was deemed to be closest to fluid intelligence. In this instance, the gains were skewed towards more educated and higher income families. Only the PIAT-math showed FE gains, which the authors believe is difficult to explain by a nutrition hypothesis. This study showed no race or sex related differences in FE gains.

2.6. Right tail gains

Only one study examined the FE in a data set that is limited to very high IQ individuals. Wai and Putallaz (2011) examined the huge (1.7 million scores) American data set of 7th grade students who took the SAT and ACT and 5th and 6th grade students who took the EXPLORE test. These tests are given to students who have scored in the top 5% for their grade on a standardized test (composite or subtest), and are part of the Duke Talented Identification Program 7th grade

search.

Flynn (1996) argued that the gains were present at all levels, but did not have data specific to the high range that is usually considered as gifted. Wai and Putallaz found the following generational IQ gains in the top 5%:

- 5.1 SAT-M
- 13.5 ACT-M
- 11.1 EXPLORE-M

The gains were concentrated on math and nonverbal subtests (see previous comments on Ang et al., 2010).

Wai and Putallaz also examined SAT-M scores of 500 and above (top 0.5%) and equivalent scores for the ACT, with the following results:

- SAT-M 1981–1985, 7.7% at or above 500
- 2006–2010, 22.7% at or above 500
- ACT-M 1990–1995, 17.7% at or above a similar level
- 2006–2010, 29.3% at or above a similar level

The obvious conclusion is that either there are a lot more truly bright children in the 2006–2010 set, or the test results are showing a significant score inflation that is not merited. They also used multigroup confirmatory factor analysis to determine whether the data sets were invariant with respect to cohort; they were not. Consequently, it can be concluded that something changed in the test construct from one cohort to the other.

[The SAT 'recentered' scores in 1995 ostensibly "as an attempt to stave off international embarrassment." Source: **https://en.wikipedia.org/wiki/SAT#1995_recentering_(raising_mean_score_back_to_500)**

Cf. The section "Secular Decline in Scholastic Achievement Scores" on page 322 in Chapter 10 of Arthur Jensen's *The g Factor*. -Ed. Note]

2.7. FE gains but without a change in inspection time

Perhaps the only study to link a biological correlate of intelligence and test scores with the FE was carried out by Nettelbeck and Wilson (2004) in Australia. In 1981, Wilson conducted a study of school grades 1 through 7, administering the Peabody Picture Vocabulary Test (PPVT) and measured inspection times (IT) for each of the participants. In 2001, the study was

replicated with virtually every parameter held constant, other than the students. The study was done in the same school, with the same grade levels, using the same PPVT and the revised PPVT-III. IT was measured with the same Gerbrands tachistoscope, under identical conditions.

The results of the study were that the students in 2001 scored essentially the same on the PPVT-III as did the students in 1981 on the PPVT. The 2001 students scored almost 5 points higher when they took the PPVT (ΔIQ = 2.5). IT measurements were the same to within the error bands. Thus, the FE was shown, but was not accompanied by improvements in IT. I asked Nettlebeck if there were any observable differences in SES or nutrition between the two groups. He said that the area served by the school was stable and that there were no observable differences in such things as nutrition or standard of living.

While IT does not correlate significantly with fluid intelligence (Burns & Nettelbeck, 2003; Burns, Nettelbeck, & Cooper, 1999), it does correlate with nonverbal IQ at about 0.50 (Deary & Stough, 1996; and others) and with Raven's matrices and performance IQ. The finding suggests that FE gains were unrelated to processing speed or other factors that explain the IT to general ability correlations.

3. Academic performance down

While IQ test scores have been rising (in some cases soaring), academic performance has done the opposite. As Jensen (1998) pointed out, when he observed that the SAT and subtests of scholastic test items have declined, real world academic performance has done the same.

Adey and Shayer (2006), of King's College London, studied the test scores of 25,000 children across both state and private schools and concluded: "The intelligence of 11-year-olds has fallen by three years' worth in the past two decades. In 1976 a third of boys and a quarter of girls scored highly in the tests overall; by 2004, the figures had plummeted to just 6% of boys and 5% of girls. These children were on average two to three years behind those who were tested in the mid-1990s."

For an assessment of how well U.S. students are doing, this URL leads to a well-written, if depressing, description of the state of teaching, education, and students: **http://www.lhup**. edu/ ~dsimanek/decline1.htm.

4. Hypothetical causes

Among the causes that have been proposed to explain the FE are these:

- Education
- Increased exposure to testing
- Exposure to artificial light
- Nutrition
- Decreased family size
- Heterosis
- More complex visual environment
- Child rearing practices
- and the use of Classical Test Theory versus Item Response Theory

4.1. Education

Since FE gains have been observed in preschool children, education is unlikely to be a cause in all data sets. As previously discussed, FE gains have usually been more pronounced on non-scholastic items, while scholastic subtests have presented lower scores at the same time and within the same tests. Direct measures of academic performance have also shown secular declines while FE gains were evident in IQ tests (Jensen, 1998). Lynn (1998) argued that the Raven tests are being inflated as a result of mathematical education; however, the relationship of simple math to increased education is a questionable factor, especially in the Colored or Standard tests (Carlson & Jensen, 1980).

Rönnlund and Nilsson (2008, 2009) examined data from the Betula prospective cohort study. This Swedish data set consists of four age-matched samples (35–80 years; N = 2,996) tested on the same battery of memory tasks. Data was taken in 1989, 1995, 1999, and 2004. A FE was found at $\Delta IQ = 1.5$ (relatively low, relative to other nations). FE gains in fluid and crystallized intelligence were approximately equal. Years of education, height (interpreted as a marker for nutrition), and sibsize [number of siblings -Ed. Note] were used as markers; together they accounted for over 94% of the time-related differences in cognitive performance. But education was a much stronger predictor than the other two items. The authors wrote: “The fact that education emerged as the strongest predictor across all cognitive measures enforces the conclusion that education may exert influence on time-related patterns on (broad) fluid (visuospatial ability, episodic memory) as well as crystallized/semantic aspects of cognition.”

4.2. Increased exposure to testing

There is little doubt that testing frequency has increased over the past years. Tuddenham listed it as one possible explanation for the secular gains he found between WW1 and WW2 cohorts. There are two mechanisms that have been proposed. Brand (1996) suggested that the use of timed tests has caused students to work faster by guessing more frequently on multiple choice tests. This largely ignored hypothesis has recently been supported by item level data (Must & Must, 2012). This finding explains other observations (lack of *g* loading in some studies and inconsistency between scoring methodologies) but does not cover all aspects of this category of causation. For example, FE gains are seen on tests that are untimed and on tests that do not use multiple choice.

Jensen (1998, p. 327) mentioned "increasing test wiseness from more frequent use of tests." His point was that frequent testing may have the same sort of impact on test scores as the increase associated with test–retest. This is the same process that is associated with learning and shows up in situations where test training has been used (as is common with the SAT). When this happens, the test *g* loading decreases and its s loading (specificity) increases.

Both Brand's and Jensen's ideas would presumably cause test scores to increase without showing gains on *g*. As will be seen later, numerous studies, but not all, have shown that FE

gains that are not *g* loaded. Flynn (2009) agreed with Jensen's comment (above), but only for the early years of testing: "The twentieth century saw us go from subjects who had never taken a standardized test to people bombarded by them, and, undoubtedly, a small portion of gains in the first half of the century was due to growing test sophistication. Since 1947, its role has been relatively modest."

4.2.1. Estonian data supports Brand's hypothesis

Brand (1996) wrote: "The correct strategy for testees is: 'When in doubt, guess.'" This hypothesis has been occasionally noted in the literature, but seldom described as a likely and significant driver of FE gains.

Item level data was preserved for the Estonian National Intelligence Test, from 1933/1936 and 2006. These data show a change in test taking strategy that is best described as increased guessing (Must & Must, 2012). The numbers of correct answers increased (SD .79), but that increase was accompanied by an increase in incorrect answers (SD .15). The number of missing answers decreased. Scores were not penalized by wrong answers, but were boosted by correct answers. The Estonian data showed relatively little guessing effect for comparisons and other simple tasks, but had a large presence on time-pressured and mentally taxing tasks (math). In the 1934–1936 tests the item level data do not suggest the guessing strategy that is apparent in the 2006 tests. It should be noted that these same data show FE gains in excess of those that can be attributed to a guessing strategy.

4.3. Nutrition and medical care

Both nutrition and medical care have improved over the past century and have been accompanied by a large number of gains that appear to be caused by these improvements: increased mean height, increased head size, faster growth, earlier maturation, etc. Lynn (2009a) argues that gains in developmental quotients (DQs — hold up head, sit up, stand, walk, jump, etc.) are indicators

of gains in IQ. DQs have gained 3.7 points per decade, while IQ gains of 3.9 points per decade have been seen in preschool children (age 4–6). Using the Griffiths Test, British children at age 6 months showed an average DQ gain of 2.8 points per decade and children, age 18 months, showed an average gain of 2.1 points per decade. Flynn (1984b) and Bocerean, Fischer, and Flieller (2003) have reported IQ gains that are similar to the DQ gains (Hanson, Smith, & Hume, 1985) for preschool children.

Lynn (2009a,b) cites various studies that show poor nutrition in the early part of the 20th century in the U.S. and Western Europe. Those indications of poor nutrition disappeared over the course of that century. Three nutrients that are known to be related to the development of intelligence are iron, folate, and iodine. Lynn (2009a) presented references showing insufficient intake of these in various countries in the early part of the 20th century. Liu et al. (2012), pointed to improvements in standard of living, nutrition, and education as possible causes for the gains in China. The studies that have shown greater FE gains in the lower part of the IQ distribution are consistent with the nutrition argument.

4.3.1. Birth weights

One factor influencing birth weight is pre-natal nutrition. Birth weight correlates positively with IQ and with DQs. Brazelton, Tronik, Lechtig, Lasky, and Klein (1977) reported that when birth weights reached 3,500 g, infants were advanced by approximately 15 DQ points at age 28 days (compared with lower birth weight babies). Low birth weights show the opposite; Drillien (1969) reported DQ score depressions of 12 points for infants with birth weights under 2,000 g, compared to those with birth weights over 2500 g (ages 6 months through 2 years). Various other studies have reported similar findings. In general, improved pre-natal nutrition increases birth weights and head size [birth weight is correlated with head size at r =0.75 (Broman, Nichols, & Kennedy, 1975).]. It is head size that is directly linked to higher cognitive performance.

[3,500 grams ~ 7.7 pounds, 2,000 grams ~ 4.4 pounds, 2,500 grams ~ 5.5 pounds -Ed. Note] *4.3.2. Height*

Lynn (2009a) attributes the change in height and in DQs as being caused by nutritional improvements. Both measures increased by about one standard deviation (SD) over 50 years. Flynn (2009) countered that gains in height have not happened at the same times as gains in IQ. This argument seems to imply a degree of data tracking, with respect to time, that is not necessary for the argument to hold (Lynn, 2009a). Height and intelligence gains for Norwegian conscripts were reported by Sundet et al. (2004) continuing until the late 1980s, when height gains ended. For the period from 1969 to 2002, the height gains were more pronounced in the upper half of the distribution, while intelligence gains were greater in the lower half.

4.3.3. Head size

Lynn (2009a) cited numerous sources that have reported head size increases of about one standard deviation over the past 50-plus years. In Britain, the head circumference of 1 year olds has increased by approximately 1.5 cm from 1930 to 1985 (Cole, 1994). Head circumference, DQs, IQs, and height, over that time span, have all shown gains of about 1 SD. Head size is an approximate measure of brain size; the two correlate at r = 0.8 (Brandt, 1978).

Jensen (1998) found that head size is mostly correlated with *g* (as opposed to group factors) and notes that the reason for the correlation is that head size is a proxy for brain size. When measured with MRI, the correlation between brain size and IQ is about 0.40 (Rushton & Ankney, 1996). Larger brain size means more neurons and is logically consistent with the correlations between head and brain measurements versus IQ.

The correlation between brain volume and IQ is presumably due to the larger number of neurons in larger brains (Rushton & Ankney, 1996), although Miller (1994) has suggested that it may be due to higher levels of myelination in larger brains. In any case, increases in brain size should be direct contributors to higher intelligence (Miller & Penke, 2007).

4.3.4. Not nutrition

- Neisser (1998) pointed out that studies of nutrition have shown that neither vitamins nor supplements have had any impact on intelligence.
- Nutrition is unlikely to have declined over the past 20 years in those countries that have a negative FE; height did not decline.
- Contrary to the intelligence gains seen in Norway, height gains from 1969 to 2002 were mostly in the upper half of the intelligence range (Sundet et al., 2004).
- With the exception of Spain, Denmark, and Norway, gains have not been frequently concentrated in the bottom half of the distribution. Flynn and Rossi-Casé (2012) argued that for all other cases, the nutrition argument is not viable.
- Mingroni (2007) argued that all postnatal environmental factors are implausible because of the high consistency of heritability estimates.
- Mingroni (2007) also contended that heterosis is a better explanation for increases in height than are nutritional and health care considerations.

4.4. Exposure to artificial light

This hypothesis is not seen often in the literature and might have been omitted in this review, except that it did not come from a weak source, but was one of the items listed by Jensen in *The g Factor*. The idea is based on the response of the pineal gland in animals to artificial light. The pineal gland appears to play a major role in sexual development, hibernation, metabolism, and seasonal breeding. Artificial light is used by poultry farmers to stimulate growth and increase their output.

There does not seem to be any data available for whether this effect happens in humans, but the speculation is that it might. There has been an obvious increase in the use of electric lighting by humans over much of the time that the FE has been observed. Besides lighting, people have been increasingly exposed to artificial light from television and computer screens, even during early childhood.

4.5. Decreasing family size

It has been known for some time that the mean IQ of families decreases as family size increases. There are two factors that contribute (presumably independently) to this effect:

- Maternal IQ correlates negatively with fertility. This is the underlying factor behind Richard Lynn's papers and book relating to global dysgenics and has been shown for numerous data sets from various countries (Lynn, 1996; Lynn & Harvey, 2008). Low IQ people statistically have more children than high IQ people. The high heritability of intelligence, therefore, is a source of dysgenic pressure. If there is a decrease in average family size (not limited to the upper end), the reduced numbers of low IQ children should produce a net increase in the mean, which would show up as a FE gain.
- Dating as far back as Sir Francis Galton, it was believed that IQ declined as a function of birth order. That belief was disputed by Rodgers, Cleveland, van den Oord, and Rowe (2000) after they examined the American NLSY data and did not find a birth order effect. This argument seemed strong and held until Bjerkedal, Kristensen, Skjeret, and Brevik (2007) published a study based on a very large data set of Norwegian conscripts, which showed the birth order effect in Norway. The mechanism of the effect has not been resolved. Hypotheses that have been advanced include prenatal gestational factors and social factors. The former seem more consistent with the general finding that social factors have little, if any effect on intelligence. Causation of the birth order effect does not matter with respect to the FE. If family size is declining in various groups, there must be a positive contribution to mean IQ due to fewer low IQ children being born.

4.6. Heterosis

Mingroni (2004, 2007) suggested that since the effects of the environment on intelligence are so small (Loehlin, Horn, & Willerman, 1989; Scarr & Weinberg, 1978), the possibility of a genetic effect should be investigated. If environmental factors were significant, between-family variance would cause MZA twins (identical, reared apart) to be less alike and siblings to be more alike.

[MZA = Monozygotic twins reared apart -Ed. Note]

Besides IQ, there have been secular trends in height, growth rate, myopia, asthma, autism, ADHD, and head circumference. It may, therefore, seem reasonable to argue that there is a global change that is affecting some or all of these factors (possibly consistent with Lynn's nutrition hypothesis). If selective breeding was involved, in order to produce the magnitudes seen in the FE, breeding would have to be restricted to only those people in the upper half of the IQ distribution (Jensen 1998, p. 327). As previously discussed, it is the bottom half that has the higher fertility.

Lynn (2009a) argued that heterosis is unlikely for three reasons:

1. There was little immigration in Europe before 1950 (the FE was present before that date).
2. The FE for IQs and DQs is just as large in Europe as in other places.
3. Studies of heterosis have shown little positive effect on IQ.

Woodley (2011) also concluded that heterosis is an unlikely cause because the FE gains are seen on the least *g* loaded components of intelligence tests [Colom, Juan-Espinosa, and Garcia (2001) reported opposite findings for Spanish standardizations of the DAT.].

[DAT = Differential Aptitude Test -Ed. Note]

Perhaps the most important consideration in determining whether there is a heterosis effect was pointed out by Mingroni: If the FE is found within-families, the cause is not genetic. Sundet, Eriksenb, Borren, and Tambs (2010) found that the FE operates within sibships. Unless this finding cannot be extended beyond Norway, the heterosis hypothesis does not look viable.

Mingroni (2007) argued in favor of a heterosis explanation from the perspective of real gains on intelligence and did not address situations, such as increased exposure to testing (Section 4.2), that show a FE, but which are inherently not Jensen effects. He also argued that increases in height were better explained by heterosis than by nutrition, but did not address that at least some of the height gains are related to leg length and are best explained by sexual selection (Jensen, 1998, p. 331).

4.7. Enriched visual environment

Greenfield (1998) and others suggested that the FE gains are caused by the ever increasing shift from verbal communication to visual and interactive media. This is seen globally in the increased presence of movies, television, photography, video games, computers, puzzles, mazes, exploded views, etc. Advertising has become ubiquitous and is saturated with images, graphs, charts, and rapid sequence visuals.

The mechanism for this hypothesis is that the shift towards visual representations removes some of the novelty from tests, especially in the culture reduced tests that have shown about double the FE gains as found in other tests. This is particularly convincing for tests such as the Raven which presents abstract figures in a matrix. Several decades ago these figures may have been more baffling than they are today.

4.8. Child rearing practices

The FE has been seen throughout the world, in both developed and undeveloped countries where child rearing practices vary greatly. It is unlikely that this hypothesis is a significant factor, not only because of the cultural variation in child rearing practices, but also because the shared environment has essentially no impact on adult intelligence (per prior discussion). To some extent, this category overlaps the increased visual environment and education. In that regard, it may contribute to the FE in some instances.

4.9. Methodological and test construct issues

As previously mentioned, ceiling effects can distort FE measurements. Other methodological issues have been found, but not fully resolved.

4.9.1. Is the FE invariant?

When researchers have tested for invariance, they have found that the data sets they were examining were not invariant (Must et al., 2009; Wai & Putallaz, 2011). Wicherts et al. (2004) did a study of five data sets to test for invariance. These included the Must et al. and Teasdale & Owen studies. Multigroup confirmatory factor analyses of these data sets showed that they were

not invariant, meaning that FE gains were not gains on the latent variables that the tests were supposed to measure. Besides providing insight as to the nature of the FE gains, the rejection of factorial invariance demonstrates that subtest score interpretations are necessarily different over time.

Flynn (2009) pointed out that cultural changes over time cause some test items to become easier because they have lost their novelty. Some words that were previously not common become more common because usage has changed. He gives several examples of this, including his frequently used example: “What do dogs and rabbits have in common?” He says that past generations would more likely focus on the use of dogs to hunt rabbits, while later generations would immediately identify that they are both mammals. This example of differential item functioning is probably responsible for at least some subtest score increases, especially in tests of similarities and vocabulary. Periodic test revisions should remove these non-*g* gains.

4.9.2. Classical Test Theory versus Item Response Theory

Beaujean and Osterlind (2008) did an analysis that is related to the Wicherts et al. analysis of invariance, which examines the underlying nature of the test itself. Most studies in the literature are based on Classical Test Theory (CTT) and present results which are not based on item level analysis. This practice hides some of the information that could be extracted from a data set. Test scores are given, but the latent constructs they are designed to measure cannot be examined. Item Response Theory (IRT), on the other hand, allows the researcher to examine the changes in underlying latent ability. Thus, CTT can show differences in scores, even when there is no change in the latent variable. An increase may be due to a general gain in real intelligence, or a decrease in the levels of difficulty of test items.

Despite its relatively infrequent use, IRT is generally considered to be the better methodology. It is particularly useful in FE studies because it reveals changes in item properties between two groups measured at different times. CTT requires groups that are being compared to have similar ability distributions, but this is not a requirement when IRT is used. In IRT, the item parameters do not depend on the ability level of the testees.

Results using CCT and IRT to measure FE gains in the American NLSY data: ● *Peabody Picture Vocabulary Test-Revised (PPVT-R)*

CCT 0.44 points per year

IRT 0.06 points per year

- Peabody Individual Achievement Test-Math (PIAT-M)

CCT 0.27 points per year

IRT 0.13 points per year

The results show that the FE essentially vanishes for the PPVT-R when IRT is used. The PIAT-M gains are cut to half using IRT. Ergo, the FE gains are a function of the methodology, leading to the concern that much of the literature has reported findings that might be quite different if IRT had been used.

Now that an item level study has been reported for the Estonian data (see Section 4.2.1), it is apparent that some of the score gains were due to increased guessing on the most complex subtests. Shiu, Beaujean, Must, te Nijenhuis, and Must (2012) reported effect sizes for the FE gains in this data set. All subtests, except computations, showed gains; the largest gain was in analogies. The research group concluded that there was some real increase in abilities (beyond the guessing related gains previously discussed).

5. Real or hollow gains?

When David Wechsler studied his WAIS, he gave the old 1953 version and the new revised 1978 version (WAIS-R) to the same group. That group averaged 103.8 on the new version and 111.3 on the old version yielding ΔIQ = 3 (Neisser, 1997).

If children of 1997 took the 1932 Stanford-Binet, 1/4 would score above IQ 130 (an increase of 10X). If children in 1932 took the 1997 test, the mean would be about 80! 1/4 would be "deficient" (Neisser, 1997).

Vroon made a similar observation about Dutch men: When scored against 1982 norms, men in 1952 would have had a mean IQ of 79 (Neisser, 1998).

Flynn initially questioned the reality that intelligence has increased:

"Has the average person in The Netherlands ever been near mental retardation?" "Does it make sense to assume that at one time almost 40% of Dutch men lacked the capacity to understand soccer, their most favored national sport?" He noted that there are not more gifted Dutch school children now and that patented inventions have shown a sharp decline.The U.S. mean in 1918 would have been 75, if scored against today's norms.If the score gains were real intelligence gains, real-life consequences would be conspicuous (Neisser, 1998). In discussing paradoxes related to the secular gains, Flynn (2009) wrote: "How can people get more intelligent and have no larger vocabularies, no larger stores of general information, no greater ability to solve arithmetical problems? …Why do we not have to make allowances for the limitations of our parents?"

5.1. Is the Flynn effect a Jensen effect?

[A Jensen effect is one that loads on *g*. It was named by Rushton.]

- Colom et al. (2001) Paper title: The secular increase in test scores is a "Jensen effect."
- Must et al. (2003) Paper title: The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect.
- Rushton and Jensen (2010): "The Flynn effect is not a Jensen effect (because it does not occur on *g*)."

5.1.1. Not a Jensen effect

In a meta-analysis of 64 test–retest studies using IQ batteries (total N = 26,990), te Nijenhuis, van Vianen, and van der Flier (2007) found a correlation between *g* loadings and score gains of −1.00. A similar finding was reported for a different meta-analysis by van Bloois, Geutjes, te Nijenhuis, and de Pater (2009). Must et al. (2003) found (in Estonia) a correlation of −0.40

between *g* and FE gains. These all show that the gains were not on *g* and were, therefore, hollow. The discussion in Section 4.2.1 shows that at least part of the Estonian gains were the result of an increased tendency to guess.

Rushton and Jensen (2010) showed that heritabilities calculated from twins also correlate with the *g* loadings, $r = 0.99$, $P < 0.001$ (for the estimated true correlation), providing biological evidence for a genetic *g*. The importance of this is that if the FE is being driven by environmental factors, it is unlikely that the gains would load on *g*. If the cause is genetic (as in the Mingroni hypothesis), the gains should show a Jensen effect.

They also pointed out that *g* loadings and inbreeding depression scores on the 11 subtests of the WISC correlate significantly positively with racial differences and significantly negatively (or not at all) with the secular gains. This is further evidence that the FE is caused by environmental factors.

Perhaps the strongest argument that the FE does not load on *g* came from Rushton (1999). He used principal components analysis to show the independence of the FE from known genetic effects.

- The IQ gains on the WISC-R and WISC-III form a cluster. This means that the secular trend is a reliable phenomenon.
- This cluster is independent of the cluster formed by racial differences, inbreeding depression scores (purely genetic), and *g* factor loadings (largely genetic). The secular increase is, therefore, unrelated to *g* and other heritable measures.

Must et al. used the Method of Correlated Vectors (see Jensen, 1998) to test the FE gains for *g* loading. Rank order correlations between the various subtests and the rank of those subtests on the *g* factor were negative and nonsignificant: $r = -.40$ (one-tailed $P = .13$). Subtests with the lowest *g* loadings showed the greatest FE gains. The authors concluded: “In Estonia, the Flynn effect is not a Jensen effect.”

5.1.2. Yes, it is a Jensen effect

Colom et al. (2001) examined two successive Spanish standardizations of the Differential Aptitude Test (DAT) battery and found gains on *g*, $r = .78$; $P < .05$. Colom: “Not a ‘Jensen effect’ is true for crystallized tests but not for fluid tests.” Using the DAT, Colom et al. showed that

subtest gains increased as their rank order of *g* loading increased [the subtests in the DAT are (in order of increasing *g* loading) numerical ability, verbal reasoning, mechanical reasoning, abstract reasoning, and spatial relations.].

5.2. Predictive bias

Jensen (1998, p. 331) stated that the definitive test of whether FE gains are hollow or not is to apply the predictive bias test. This means that two points in time would be compared on the basis of an external criterion (real world measurement, such as school grades). If the gains are hollow, the later time point would show underprediction, relative to the earlier time. This assumes that the later test has not been renormed. In actual practice tests are periodically renormed so that the

mean remains at 100. The result of this recentering is that the tests maintain their predictive validity, indicating that the FE gains are indeed hollow.

[Editor's Note: See discussion above about SAT recentering, section 2.6] **6. Which explanations work?**

Most of the mechanisms that have been proposed as causes of the FE are plausible under some circumstances. Even when one is ruled out by a specific study, it may apply elsewhere. As has been shown in the foregoing material, the most consistent aspect of the FE is that it is inconsistent from one time or place to another. Sometimes the gains have been mostly in abstract reasoning (as in the U.S.), but elsewhere the gains have been strongly tilted towards scholastic subtests (Estonia). Gains have been strong, weak, flat, or have reversed, even within the same country when measured at different times — Norway and Denmark (Sundet et al., 2004; Teasdale & Owen, 2008).

Finally, there are the issues of non-invariance and of methodological inconsistency when IRT is used instead of CTT. The instances in which confirmatory factor analysis has failed to show invariance (every case so far) tell us that the meaning of IQ tests is not constant over time. The reduction in FE magnitude (to near zero in some cases) when IRT is applied suggests that the test vehicle is contributing 50 to 100% of the gains and that those gains are methodological artifacts and carry no *g* loading. For example, the FE gains due to guessing (Estonia) were not resolved by CTT because the successful strategy was not apparent at the subtest level.

6.1. Real or hollow?

Most of the tests for *g* loading have shown little or no *g* saturation. The majority of researchers who have addressed the issue have argued that the gains are hollow, with the exception of Lynn and Colom, both of whom have made strong arguments that there is at least some genuine gain in intelligence. This inconsistency may be due in part to different data sets and may be due in part to CTT methods. It is likely that most of the FE gains that have been reported are hollow. If this were not true, renorming would cause predictive validity to change, but there are no reports that this has happened.

7. Can the Flynn effect be modeled?

Most studies of the FE have attempted to apply a single explanation, such as heterosis, or a narrow category of causation, such as nutrition/health care. This overview, however, strongly suggests that multiple causes are acting, and that the mix of causes varies over time and from one place to another. Flynn and Rossi-Casé (2012) agree: "Even in developed nations, the notion that the Flynn effect will have identical causes should be banished from the literature."

A quantitative model of causation is beyond present understanding, but a qualitative model can be constructed, such that the most likely active components can be identified. Two approaches to this follow.

7.1. A life history model

Woodley (2012) presented a model in which a large number of FE causes (as discussed here) are assumed to vary as a group. His model assumes that the FE gains are unrelated to *g* and are the

result of a shift in life history from fast to slow. A fast life history is taken to be the set of tradeoffs that are associated with relatively high fertility and lower parental investment in offspring, as described by Rushton (1985) in his Differential K Theory; slow life history is the opposite (lower fertility and more parental investment). Woodley describes his model as a cognitive differentiation–integration effort (CD–IE) hypothesis.

- Cognitive integration effort (CIE) – a strengthening of the manifold via the investment of bioenergetic resources – fast life history.
- Cognitive differentiation effort (CDE) – a weakening of the manifold via the unequal investment of resources into individual abilities – slow life history.

If it happens that a given population is moving from a fast towards a slow life history, multiple environmental factors can be expected to move in the direction that would cause a secular rise in test scores: fertility, education, pathogen stress, and nutrition.

7.2. Independent Drivers model

The Woodley model, described above, focuses on a latent variable, such that variations in that variable contribute to the FE by means of the causes that are assumed to increase or decrease together. An alternative model assumes that the various FE drivers act independently, may combine in any combination, and may include negative driver components. The causes that are present in a given data set over an observation period are difficult to quantify, but can be estimated on a limited scale, such as high, medium, and low, with the expectation that their contributions to FE gains will be larger or smaller, depending on the strength of the driver.

Each driver is assumed to exert a FE influence as a function of how much contribution potential remains in association with that driver. For example, the reduction in family size is likely to initially contribute more to a study group that has had high fertility and is moving in the direction of smaller families. As the process continues, diminishing FE gains will be seen as the maximum total effect is used up. The path may appear to be somewhat linear over a short time period, but it must approach an asymptote. The gain for any given driver should follow a relationship that is similar to

$$FEG_i = FEM_i\ (t) / (t+k_i)$$

where FEG_i is the FE gain due to driver i; FEM_i is the maximum FE gain that can be contributed by driver i; t is the time in years; and k_i is a constant for driver i. Multiple drivers would be additive, but each will have its own maximum contribution and constant.

The shape illustrated in Fig. 1 is consistent with the gains (general shape) shown by the Raven's Progressive Matrices in Britain (Hiscock, 2007).

7.2.1. Reversals

Reversals may occur either as the sum of positive drivers decreases to less than the sum of negative drivers, or the positive drivers reverse direction. A lack of FE push might result in a reversal due to an existing negative cause, such as an underlying dysgenic trend or the decline in educational participation. The net FE gain (or loss) may contain negative factors that are not

evident in the data, because the result is a positive FE. Thus, the positive drivers need only reach saturation for a reversal to appear (assuming the presence of one or more negative drivers).

It is possible that some of the drivers that have been discussed could reverse direction and directly cause a FE decline. For example, nutritional factors may change and become negative due to the introduction of harmful chemicals into diets or the living environment; health care standards could deteriorate; family sizes could reverse direction, at least for a segment of a population.

7.2.2. FE Drivers

Group and environmental characteristics over the time period ΔT	FE driver
Many school years completed	Education
Qualitatively improved education	
Higher scores on scholastic tests	
Score gains in preschool children	Not education, but possibly
More testing in primary and secondary schools	nutrition, healthcare, etc.
Increased use of tests for college level selection	Increased exposure to testing
Recent electriciation, as might be seen in remote areas	Exposure to artificial light
Increased availability of television	
Growth of personal computers in homes and schools	
Increased pediatric care	Nutrition and healthcare
Diet improvements of critical nutrients	
Mean increases in weight, head size, or birth weight	
Accelerated childhood development	
Lower fertility for low SES levels	Decreased family size
Increased availability of television	
More complex visual environment	
Growth in personal computers in homes and schools	
Increased visual complexity of school textbooks	
Advertising growth, accompanied by charts, symbols, etc.	
Measured increase in mean *g*	Nutrition and healthcare
Change in breeding pattern from isolated groups to	
breeding among groups, not accompanied by	Decreased family size within-family FE
Heterosis	

For a given data set, the presence of items from the first column implies a cause from the second column. For example, Must and Must (2009) reported a height increase (in Estonia) of 2.9 SD over approximately 2 centuries. At the beginning of the 20th century, the diet was primarily bread and herring. From 1925 to 1958 there was a shift from vegetarian foods to meats. This pattern signals that the nutritional FE driver was active during and after the dietary change. FE gains were seen in scholastic performance and reasoning, suggesting that education was also a

factor. The general increase in prosperity of the country may also signal matches for other changes (first column), such as decreased family size.

In some situations, the Independent Drivers model could reduce to the Woodley model, but in situations where the effect can only be linked to one or two drivers, this model is accommodating. In any situation where a gain in *g* is seen, the Woodley model would not apply, but this model identifies nutrition, health care, and heterosis as possible *g* loaded drivers.

8. Summary

- The FE exists between birth cohorts.
- It has been found within sibships.
- It sometimes appears early in life (before school age).
- There are presumably multiple causes.
- The gains are often hollow (not Jensen effects) but some gains appear to be on *g*.
- There are methodological issues to be resolved which may be a cause of some of the gains.
- The FE is not invariant over time.

9. Recommendations

Despite the huge mass of papers, the FE remains enigmatic. Part of the problem is the complication of what strongly appears to be varying combinations of multiple drivers; individual studies cannot be consistently compared. But the concern that deserves particular attention is that methodological issues appear to be confounded with real world causes. Perhaps ways can be found to examine more data sets with IRT. It would be very helpful to know how much of the various FE gains are the result of CTT methodology. The findings of non-invariance presumably mean that some FE gains are attributable to test revisions and to cultural shifts. A better grasp of the categories of test items that are causing non-invariance may enable test designers to reduce or eliminate these test-specific items.

Fig 1. Flynn effect gains for a single driver. In the illustration the maximum contribution for the driver is shown as 3 IQ points and the value of k is set at 2.

[Editor's Note: X axis reads "Time, years" and Y axis reads "IQ point gains from one driver"]

Some direct connections between environmental conditions and the FE have been identified, such as those in Estonia (dietary changes, family size reductions, and educational improvements). These point to causes for a single country, but cannot be generalized. Future researchers should be encouraged to examine national data sets from health and social service agencies to identify sharp changes that correspond to FE rate changes. Some of this has already been done by Lynn, but there may be additional factors that have not yet surfaced. In the U.S. the National Institute of Health and the Food and Drug Administration are probable data sources. Other environmental factors that might be worth examining for coincidence with FE rate changes: the introduction of radio, television, computers, the Internet, and cell phones, etc.

Educational policies and numbers of graduates might be considered as well, despite declines in academic performance, there may still be FE drivers associated with formal or informal education.

Finally, it would be helpful to perform studies of biological parameters that relate to intelligence. There is the IT study by Nettelbeck & Wilson, but little else in this category. The question to answer is whether other biological measurements (RT, brain pH, nerve conduction velocity, pitch discrimination, EEG latencies, glucose uptake rates, etc.) remain stable over decades, or do they vary in the direction that would be predicted by an increase in intelligence?

Acknowledgment

I would like to thank James Thompson for his constructive comments on this manuscript.

Bibliography

Adey, P., & Shayer, M. (2006). Cited by Guardian Co Uk, (available at **http://**

education.guardian.co.uk/schools/story/0,1693061,00.html).

Ang, S., Rodgers, J., & Wänström, L. (2010). The Flynn effect within subgroups in the U.S.: Gender, race, income, education, and urbanization differences in the NLSY-Children data. Intelligence, 38–4, 367–384.

Bayley, N. (1993). Bayley scales of infant development. San Antonio, TX: Psychological Corporation.

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 Children and Young Adults Data. Intelligence, 36(5), 455–463.

Bjerkedal, T., Kristensen, P., Skjeret, G. A., & Brevik, J. I. (2007). Intelligence test scores and birth order among young Norwegian men (conscripts) analyzed within and between families. Intelligence, 35–5, 503–514.

Black, M. M., Hess, C. R., & Berenson-Howard, J. (2000). Toddlers from low-income families have below normal mental, motor and behavior scores on the revised Bayley Scales. Journal of Applied Developmental Psychology, 21, 655–666.

Bocerean, C., Fischer, J. -P., & Flieller, A. (2003). Long term comparison (1921–2001) of numerical knowledge in 3 to five and a half year old children. European Journal of Psychology of Education, 18, 405–424.

Brand, C. (1996). The g factor: General intelligence and its implications. Chichester, England: Wiley.

Brandt, I. (1978). Growth dynamics of low-birth weight infants with emphasis on the perinatal period. In F. Falkner, & J. M. Tanner (Eds.), Human growth, Vol. 2. (pp. 557–617)New York: Plenum.

Brazelton, T. B., Tronik, E., Lechtig, A., Lasky, R. E., & Klein, R. E. (1977). The behavior of nutritionally deprived Guatemalan infants. Developmental Medicine and Child Neurology, 19, 364–372.

Broman, S. H., Nichols, P. L., & Kennedy, W. A. (1975). Preschool IQ: Prenatal and developmental correlates. Hillsale, NJ: Wiley.

Burns, N., & Nettelbeck, T. (2003). Inspection time in the structure of cognitive abilities: Where does IT fit? Intelligence, 31, 237–255.

Burns, N. R., Nettelbeck, T., & Cooper, C. J. (1999). Inspection time correlates with general speed of processing but not with fluid ability. Intelligence, 27, 37–44.

Campbell, S. K., Siegel, E., Parr, C. A., & Ramey, C. T. (1986). Evidence for the need to renorm the Bayley Scales of Infant Development based on the performance of a population-based sample of 12 month old infants. Topics in Early Childhood Education, 6, 83–96.

Carlson, J. S., & Jensen, C. M. (1980). The factorial structure of the Raven Coloured Progressive Matrices Test: A reanalysis. Educational and Psycho- logical Measurement, 40, 1111–1116.

Cole, T. J. (1994). Growth charts for both cross-sectional and longitudinal data. Statistics in Medicine, 13, 2477–2492.

Colom, R., Andre's Pueyo, A., & Juan-Espinosa, M. (1998). Generational IQ gains: Spanish data. Personality and Individual Differences, 25(5), 927–935.

Colom, R., Flores-Mendoza, C. E., Francisco, J., & Abad, F. J. (2007). Generational changes on the draw-a-man test: A comparison of Brazilian urban and rural children tested in 1930, 2002 AND 2004. Journal of Biosocial Science, 39, 79–89.

Colom, R., Juan-Espinosa, M., & Garcia, L. F. (2001). The secular increase in test scores is a "Jensen effect.". Personality and Individual Differences, 30, 553–559.

Colom, R., Lluis-Font, J. M., & Andres-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. Intelligence, 33, 83–91.

Cotton, S. M., Kiely, P. M., Crewther, D. P., Thomson, B., Laycock, R., & Crewther, S. G. (2005). A normative and reliability study for the Raven's Colored Progressive Matrices for primary school aged children in Australia. Personality and Individual Differences, 39, 647–660.

Deary, I. J., & Stough, C. (1996). Inspection time and intelligence: Achievements, prospects and problems. American Psychologist, 51, 599–608.

Drillien, C. M. (1969). School disposal and performance for children of different birthweight born 1953–1960. Archives of Diseases in Childhood, 44, 562–570.

Flynn, J. R. (1984a). IQ gains and the Binet decrements. Journal of Educational Measurement, 21, 283–290. Flynn, J. R. (1984b). The mean IQ of Americans: Massive gains 1932 to 1978. Psychological Bulletin, 95, 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. Psychological Bulletin, 101, 171–191.

Flynn, J. R. (1996). What environmental factors affect intelligence: The relevance of IQ gains over time. In D. Detterman (Ed.), Current topics in human intelligence, vol. 5: The environment. Norwood, NJ: Ablex.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. American Psychologist, 54(1), 5–20.

Flynn, J. R. (2009). What is intelligence? Beyond the Flynn effect. Cambridge: Cambridge University Press.

Flynn, J. R., & Rossi-Casé, L. (2012). IQ gains in Argentina between 1964 and 1998. Intelligence, 40, 145–150.

Greenfield, P. M. (1998). The cultural evolution of IQ. In U. Neisser (Ed.), The rising curve: Long-term gains in IQ and related measures (pp. 81–123). Washington, DC: American Psychological Association.

Hanson, R., Smith, J. A., & Hume, W. (1985). Achievements of infants on items of the Griffiths scales: 1980 compared with 1950. Child: Care, Health and Development, 11, 91–104.

Herrnstein, R. J., & Murray, C. (1994). The bell curve: Intelligence and class structure in American life. New York: Free Press.

Hiscock, M. (2007). The Flynn effect and its relevance to neuropsychology. Journal of Clinical and Experimental Neuropsychology, 29(5), 514–529.

Jensen, A. R. (1998). The g factor: The science of mental ability. Westport, CT: Praeger.

Kagitcibasi, C., & Biricik, D. (2011). Generational gains on the draw-a-person IQ scores: A three-decade comparison from Turkey. Intelligence, 39, 351–356.

Kanaya, T., Ceci, S. J., & Scullin, M. H. (2005). Age differences within secular IQ trends: An individual growth modeling approach. Intelligence, 33, 613–621.

Liu, J., Yang, H., Li, L., Chen, T., & Lynn, R. (2012). An increase of intelligence measured by the WPPSI in China, 1984–2006. Intelligence, 40, 139–144.

Loehlin, J. C., Horn, J. M., & Willerman, L. (1989). Modeling IQ change: Evidence from the Texas Adoption Project. Child Development, 60, 993–1004.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. Nature, 297, 222–223. Lynn, R. (1996). Dysgenics: Genetic deterioration in modern populations. Praeger Publishers.

Lynn, R. (1998). In support of nutrition theory. In U. Neisser (Ed.), The rising curve. Washington, DC: American Psychological Association.

Lynn, R. (2009a). What has caused the Flynn effect? Secular increases in the development quotients of infants. Intelligence, 37(2009a), 16–24.

Lynn, R. (2009b). Fluid intelligence but not vocabulary has increased in Britain, 1979–2008. Intelligence, 37, 249–255.

Lynn, R., & Hampson, S. (1986). The rise of national intelligence. Evidence from Britain, Japan and the United States. Personality and Individual Differences, 7, 23–32.

Lynn, R., & Harvey, J. (2008). The decline of the world's IQ. Intelligence, 36, 112–120.

Miller, E. M. (1994). Intelligence and brain myelination: A hypothesis. Personality and Individual Differences, 17, 803–832.

Miller, G. F., & Penke, L. (2007). The evolution of human intelligence and the coefficient of additive genetic variance in human brain size. Intelligence, 35, 97–114.

Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. Intelligence, 32, 65–83.

Mingroni, M. A. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. Psychological Review, 114, 806–829.

Must, O., & Must, A. (December 19). The biological correlates of the Flynn effect in Estonia. Paper presented at the 10th Annual Meeting of the International Society for Intelligence Research, Madrid, Spain.

Must, O., & Must, A. (December 13). Test-taking patterns have changed over time. Paper presented at the 13th Annual Meeting of the International Society for Intelligence Research, San Antonio, Texas.

Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. Intelligence, 31, 461–471.

Must, O., te Nijenhuis, J., Must, A., & van Vianen, A. E. M. (2009). Comparability of IQ scores over time. Intelligence, 37, 25–33.

Neisser, U. (September–October). Rising scores on intelligence tests. American Scientist.

Neisser, U. (1998). The rising curve. Washington: American Psychological Association (7).

Nettelbeck, T., & Wilson, C. (2004). The Flynn effect: Smarter not faster. Intelligence, 32, 85–93.

Rodgers, J. L., Cleveland, H. H., van den Oord, E., & Rowe, D. C. (2000). Resolving the debate over birth order, family size, and intelligence. American Psychologist, 55, 599–612.

Rönnlund, M., & Nilsson, L. -G. (2008). The magnitude, generality, and determinants of Flynn effects on forms of declarative memory and visuospatial ability: Time-sequential analyses of data from a Swedish cohort study. Intelligence, 36, 192–209.

Rönnlund, M., & Nilsson, L. -G. (2009). Flynn effects on sub-factors of episodic and semantic memory: Parallel gains over time and the same set of determining factors. Neuropsychologia, 47, 2174–2180.

Rushton, J. P. (1985). Differential K theory: The sociobiology of individual and group differences. Personality and Individual Differences, 6, 441–452.

Rushton, J. P. (1999). Secular gains in IQ not related to the g factor and inbreeding depression — Unlike Black–White differences: A reply to Flynn. Personality and Individual Differences, 26, 381–389.

Rushton, J. P., & Ankney, C. D. (1996). Brain size and cognitive ability: Correlations with age, sex, social class, and race. Psychonomic Bulletin & Review, 3(1), 21–36.

Rushton, J. P., & Jensen, A. R. (2010). The rise and fall of the Flynn effect as a reason to expect a narrowing of the Black–White IQ gap. Intelligence, 38, 213–219.

Scarr, S., & Weinberg, R. A. (1978). The influence of "family background" in intellectual attainment. American Sociological Review, 43, 674–692. Shiu, W., Beaujean, A. A., Must, O., te Nijenhuis, J., & Must, A. (December 13). Item-level examination of the Flynn effect. Paper presented at the 13th Annual Meeting of the International Society for Intelligence Research, San Antonio, Texas.

Smith, S. (1942). Language and nonverbal test performance of racial groups in Honolulu before and after a 14-year interval. The Journal of General Psychology, 26, 51–92.

Sundet, J. M., Barlaug, D. G., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. Intelligence, 33, 349–362.

Sundet, J. M., Eriksenb, W., Borren, I., & Tambs, K. (2010). The Flynn effect in sibships: Investigating the role of age differences between siblings. Intelligence, 38–1, 38–44.

Tasbihsazan, R., Nettlebeck, T., & Kirby, N. (1997). Increasing mental development index in Australian children: A comparative study of two versions of the Bayley Mental Scale. Australian Psychologist, 32, 120–125.

te Nijenhuis, J. T., Cho, S. H., Murphy, R., & Lee, K. H. (2012). The Flynn effect in Korea: Large gains. Personality and Individual Differences, 53, 147–151.

te Nijenhuis, J. T., Murphy, R., & van Eeden, R. (2011). The Flynn effect in South Africa. Intelligence, 39(2011), 456–467.

te Nijenhuis, J. T., van Vianen, A., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. Intelligence, 35, 283–300.

Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. Intelligence, 13, 255–262.

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn effect. Intelligence, 36, 121–126. Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. American Psychologist, 3, 54–56.

van Bloois, R. M., Geutjes, L. -L., te Nijenhuis, J., & de Pater, I. E. (December 19). g loadings and their true score correlations with heritability coefficients, giftedness, and mental retardation:

Three psychometric meta-analyses. Paper presented at the Symposium on Group Differences, 10th Annual Meeting of the International Society for Intelligence Research, Madrid, Spain.

Wai, J., & Putallaz, M. (2011). The Flynn effect puzzle: A 30-year examination from the right tail of the ability distribution provides some missing pieces. Intelligence, 39, 443–455.

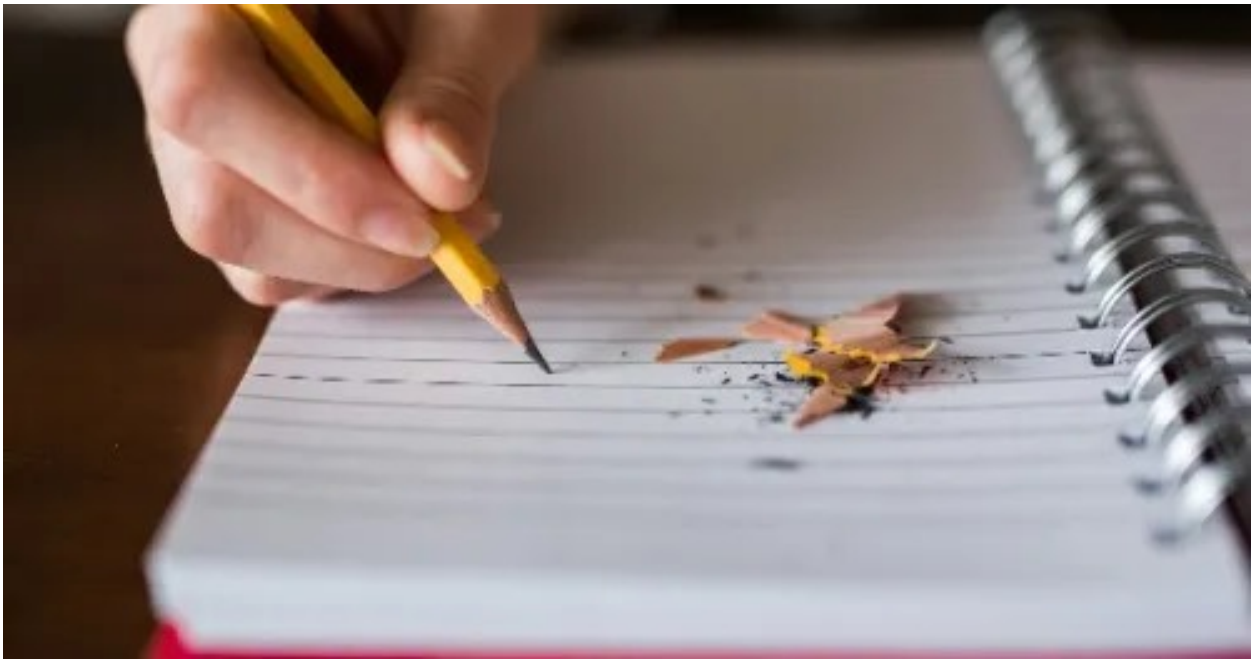
Wicherts, J. M., Dolan, C. V., Hessen, D., Oosterveld, P., Baal, G. C. M., van Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. Intelligence, 32, 509–537.

Woodley, M. A. (2011). Heterosis doesn't cause the Flynn effect: A critical examination of Mingroni (2007). Psychological Review, 118(4), 689–693.

Woodley, M. A. (2012). A life history model of the Lynn–Flynn effect. Personality and Individual Differences, 53(2), 152–156.

Chris Cole, Richard May, Rick Rosner on Debunking I.Q. Scores

2024-08-15



Chris Cole is a longstanding member of the Mega Society. **Richard May** is a longstanding member of the Mega Society and Co-Editor of *Noesis: The Journal of the Mega Society*. Alternatively: "Richard May ("May-Tzu"/"MayTzu"/"Mayzi") is a Member of the Mega Society based on a qualifying score on the Mega Test (before 1995) prior to the compromise of the Mega Test and Co-Editor of *Noesis: The Journal of the Mega Society*. In self-description, May states: "Not even forgotten in the cosmic microwave background (CMB), I'm an Amish yuppie, born near the rarified regions of Laputa, then and often, above suburban Boston. I've done occasional consulting and frequent Sisyphean shlepping. Kafka and Munch have been my therapists and allies. Occasionally I've strived to descend from the mists to attain the mythic orientation known as having one's feet upon the Earth. An ailurophile and a cerebrotonic ectomorph, I write for beings which do not, and never will, exist—writings for no one. I've been awarded an M.A. degree, mirabile dictu, in the humanities/philosophy, and U.S. patent for a board game of possible interest to extraterrestrials. I'm a member of the Mega Society, the Omega Society and formerly of Mensa. I'm the founder of the Exa Society, the transfinite Aleph-3 Society and of the renowned Laputans Manqué. I'm a biographee in Who's Who in the Brane World. My interests include the realization of the idea of humans as incomplete beings with the capacity to complete their own evolution by effecting a change in their being and consciousness. In a moment of presence to myself in inner silence, when I see Richard May's non-being, 'I' am. You can meet me if you go to an empty room." Some other resources include *Stains Upon the Silence: something for no one*, *McGinnis Genealogy of Crown Point, New York: Hiram Porter McGinnis*, Swines List, Solipsist Soliloquies, Board Game, Lulu blog, *Memoir of a Non-Irish Non-Jew*, and May-Tzu's posterous." **Rick Rosner** is a longstanding member of the Mega Society and a former editor of *Noesis: The Journal of the Mega Society*. Alternatively:

"According to some semi-reputable sources listed here, Rick G. Rosner may have among America's, North America's, and the world's highest measured IQs at or above 190 (S.D. 15)/196 (S.D. 16) based on several high range test performances created by Christopher Harding, Jason Betts, Paul Cooijmans, and Ronald Hoeflin. He earned 12 years of college credit in less than a year and graduated with the equivalent of 8 majors. He has received 8 *Writers Guild Awards*nominations, winning one and an *Emmy* nomination, and was named *2013 North American Genius of the Year* by The World Genius Directory. He has written for *Remote Control*, *Crank Yankers*, *The Man Show*, *The Emmys*, *The Grammys*, and *Jimmy Kimmel Live!*. He worked as a bouncer, a nude art model, a roller-skating waiter, and a stripper. In a television commercial, Domino's Pizza named him the "World's Smartest Man." The commercial was taken off the air after Subway sandwiches sent a cease-and-desist letter. (The commercial dramatized the results of a taste test in which Domino's sandwiches were preferred over Subway's sandwiches 2 to 1, but Subway and its lawyers claimed the taste test methodology was biased and flawed.) He was named "Best Bouncer" in the Denver Area by *Westwood Magazine*. Rosner spent some of the late Disco Era as an undercover high school student. In addition, he spent 25 years as a bar bouncer and fake ID-catcher, and 25+ years as a stripper, and nearly 30 years as a writer for more than 2,500 hours of network television. Errol Morris profiled Rosner in the interview series *First Person*. He came in second (lost) on *Jeopardy! and* sued *Who Wants to Be a Millionaire* over a flawed question and lost the lawsuit. He won one game and lost one game on *Are You Smarter Than a Drunk Person?*. (He was drunk.) He has spent 40+ years working on a semi-time-invariant version of Big Bang Theory. Currently, Rosner sits tweeting in a bathrobe (winter) or a towel (summer). He lives in Los Angeles, California with his wife and two dogs. He and his wife have a daughter. You can send him money or questions or just give him shit on Twitter, or find him on LinkedIn. He has a crappy little show on PodTV." They discuss: I.Q.; fake I.Q. and real I.Q.; more reliable and valid I.Q. ranges; robust, legitimate tests; the status of measuring I.Q. scores above 4-sigma; major warning signs of something awry; the minor, or subtle, warning signs; 4 standard deviations above the norm; the successes and failures of the Mega Test, the Ultra Test, the Power Test, and the Titan Test; 4 and 5 sigma above the norm; the principal design of the Adaptive Test; other extraordinary high-I.Q. societies; associative horizon; the Mega Test; the claims about the Mega Test; legitimate testing; extrapolations well beyond the norms of the mainstream tests; the motivation behind making claims well beyond the norms of the most used mainstream I.Q. tests; the more egregious I.Q. claims in 20th century; the big lessons in debunking phony I.Q. claims; fraudulent activity; messianic posing; criminal behaviour; the three interpenetrating cubes problem; above 4 standard deviations above the norm; the hardest IQ test; and IQ.

Scott Douglas Jacobsen: Today, as this is a group discussion with three longstanding members of the Mega Society, the focus is Intelligence Quotient or I.Q., particularly debunking claims. What is I.Q. truly a measure of, at this point?

Chris Cole: I.Q. is an attempt to measure general intelligence, which is analogous to the power of a computer. There is an enormous literature on this subject. I'm going to take it as a given. It will be embarrassing if when we understand more about how the mind works it turns out to be a chimera.

Richard May: ‘g’, the general factor of intelligence, i.e., cognitive ability.

Rick Rosner: IQ as measured by a high-end test is somewhat different from IQ as measured by a regular range usually group-administered test. Regular range tests measure intelligence, the ability to focus for 45 minutes, and cultural literacy.

High-end tests can measure obsessiveness and attention to detail, a love of puzzle-solving, and in some cases desperation for validation.

Intelligence has changed over the past 20 years to include skill at using tech to get answers.

Jacobsen: What differentiates a fake I.Q. score claim from a real one, e.g., signals of a fraud or claims far above the norms of a test, etc.?

Cole: Since it is difficult to define, it is difficult to measure. There is a desire to claim intelligence which creates a motivation for “vanity” tests. In science we try to overcome such tendencies using experiments to disprove theories. It is a sign of trouble if a test is not carefully normed.

May: You can perhaps find examples on Facebook and the social media generally.

Rosner: Concerted efforts to lie are fairly rare – claiming a high IQ is not very helpful in life and may even hurt – there’s Stephen Hawking’s quote that “People who brag about their IQ are losers.” There are casual claims – BSers at parties, movie stars trying to seem smart. Geena Davis’s PR team used to mention that she’s Mensa. Sharon Stone is said to have a 150 IQ. James Woods 180. And these might be legit. But that’s to address a specific issue of not being considered a bimbo.

One big tell for IQ fraud is people claiming to have completed and gotten a high score on the Mega or Titan in 10 or 12 hours. Back in 1985, I spent more than 100 hours on the Mega. Now with the internet (and coding skills which I don’t have), I could’ve cut that time by 80%. But the internet has also invalidated the Mega – not only with all of the answers floating around out there but also with instantly solving the verbal analogies just by plugging them into Google.

Jacobsen: What ranges for I.Q. scores have the highest reliability and validity, typically?

Cole: The Langdon and Hoeflin tests are on the cutting edge of reliability and validity. The Mega Test, for example, has been normed several different ways. A group of us are working on a new test that is cheat resistant.

May: Scores with the highest reliability and validity are those closest to the mean on standard IQ tests. Hoeflin and Langdon’s tests are untimed power tests more suitable for measuring above average intelligence.

Jacobsen: What tests are considered the most robust, legitimate?

Cole: We have a problem now that several of the most carefully normed, such as the Langdon Adult Intelligence Test, the Mega Test, the Titan Test, the Ultra Test, and the Power Test have been spoiled.

May: Those of Hoeflin, Langdon and Wechsler.

Rosner: Hoeflin's tests have been the most thoroughly revised and normed. His Mega Test was normed on more than 4,000 test takers. His test items are excellent. But his tests have been voided by the internet – too many easily found answers. The Mega was published in Omni magazine in 1985, I think, a decade before most people had the internet. You had to use actual physical dictionaries.

Today, I think Paul Cooijmans' tests are the most legit high-end tests. Paul takes pleasure in bursting the bubbles of people who claim high IQs by offering stringent scoring and norming. Doing well on his tests takes much time and what he calls "associative horizon" – being able to come up with dozens of ideas to crack a tough item.

Jacobsen: What is the status of measuring I.Q. scores above 4-sigma – experimental high-range testing, in other words?

Cole: The Adaptive Test, which is a work in progress, is the cutting edge. Contact me if you want to work on it. [Ed. chris@questrel.com.]

May: Apparently measurement at the far-right tail of intelligence has improved astronomically. I mistakenly thought that determining and measuring IQ was quite difficult even at the 4 sigma level. The Mega Society used to have a statement either at the beginning of Noesis or on our website or both, I think, indicating that we attempted to select members at the 4.75 sigma level, but selecting this rarity was experimental and quite difficult for many reasons. (Not exact wording.)

Today there is an IQ group which has apparently identified the 3 most intelligent individuals on planet Earth! This is quite an achievement in my view.

Since it is well known that the actual distribution of IQ-scores at the far-right tail does not conform to a Gaussian distribution, one has to assume that even if the ceiling of the IQ tests employed was sufficient (not exceeding that intended by the test developers) and the intercorrelation of the various tests at the highest levels was known and that the correct Kuder-Richardson (?) formulas were applied to concatenate the valid IQ scores, that the entire population of planet Earth was actually tested by or on behalf of this group. Since various planetary subgroups of different sizes could have differing means, standard deviations and distribution shapes, a weighted average would need to be taken in order to determine the statistical properties of the global IQ distribution for planet Earth.

This is an unparalleled achievement in psychometric history. I personally don't know anyone tested for this project in order to determine the actual shape of the global distribution of IQ-scores at the far-right tail, but I assume this is just a minor sampling error. Presumably you and your friends and neighbors have all been tested. Since the three most intelligent individuals on planet Earth have now been identified in fact, the correct protocols were undoubtedly used. If only Lewis Terman were alive now! — LINK here.

Jacobsen: You have all been around the block. Your membership in the Mega Society has spanned decades. So, you've seen controversies, failed high-I.Q. societies, and proclamations to this-or-that I.Q., even individuals who spun off into fraudulent activities, messianic posing, and

criminal behaviour. As a note on collectives of high-I.Q. people, when it comes to claimed high-I.Q. societies, what are the major warning signs of something awry, not quite right, with it?

May: The major warning signs of statistical and psychometric incompetence, fraud, or madness are usually quite subtle. Please see below.

Rosner: You got to start with the disclaimer that most people in high-IQ societies are well-behaved relatively normal people who like taking tests and solving puzzles, and there are only a few lunatics. And because the ones I belong to don't get together very often, you don't have a chance to see any warning signs developing.

Although, in the case of one guy from many years ago, you could see a guy who was kind of being physically dominant and, I guess, mentally dominant getting increasingly frustrated that people didn't understand him or believe his theories. So, it was just an increasing belligerence or pre-belligerence.

I guess, a skosh of megalomania.

Cole: The major warning signs are the ones you list: fraudulent activity, messianic posing, and criminal behavior.

Jacobsen: Following from the previous question, what are the minor, or subtle, warning signs?

May: I get slightly suspicious if someone comes up with the most brilliant Theory of Everything ever, explained in a newly invented language of neologisms, which only the inventor of the theory himself can understand, especially if the theory makes no falsifiable predictions and none of those few who claim to understand the theory can explain it in their own words. I'm also slightly suspicious of, e.g., taxi cab drivers or barbers, who have conclusively proved Einstein's theory of special and general relativity wrong.

If someone claims to be the most intelligent person in the history our solar system or to be the actual God of the Bible, then this level of measured intelligence may be beyond the current development of psychometric science, even with the Flynn effect. I'm probably too skeptical sometimes.

Also, branding of one's associates by high-IQ types is often unnecessary in my view.

Rosner: Again, I don't hang. I have no basis or nothing to talk about regarding this. It is not like I was living with a high-IQ person who slowly went crazy, besides myself. Really, in the last few years, I've gotten less crazy, more lazy. Lazy has replaced crazy.

Cole: The minor warning signs are incredible IQ claims. As a rule of thumb anything above five sigma is not credible as is anything that has not been normed using regular statistical methods.

Jacobsen: Why is 4 standard deviations above the norm (e.g., mean 100, S.D. 15, I.Q. 160) such a difficult barrier to break in finding highly intelligent individuals?

May: Almost no one in the alleged "real world" is interested in measuring intelligence beyond the 4 sigma level. Where would you find a large sample of individuals beyond the top 1-per-30,000 level of intelligence to study? This level of intelligence is not a target level for standard IQ tests developed by psychologists. Why should it be? Which professions require IQs

beyond the 4 sigma level? Even Nobels in physics probably depend more upon a mathematical ability sub-factor of general intelligence than upon super-high IQ per se. Two physics Nobel laureates didn't qualify for inclusion in Lewis Terman's study of the intellectually gifted, because their IQs were not sufficiently high! In addition Nature may sometimes not be 'politically correct'. What if cognitive differences were discovered among various human sub-groups? For example, what if a growing number of trans-species individuals, who identify as advanced AI units, were found to be better at arithmetic addition?

Rosner: Several reasons, one, there aren't that many people. 4-sigma level is one person in 30,000. Although, in real terms, it's less rare than that because the average IQ of people on the street is like 105 or 110. The people with IQs of 35 are institutionalized. You don't see them around. It's rare. That's one problem.

Problem two, it is hard to test. All the good high-end tests take dozens of hours to do well on. Thing two-and-a-half, many people who might score well on them might be successful and may not want to waste their time putting in 40 or 50 hours in something that doesn't compensate them.

They could be trading stocks or coding or doing business deals or getting laid. None of which taking an IQ test helps.

Cole: High range tests require high range questions which are hard to create. Plus there is not much of a market.

Jacobsen: What have been the successes and failures of the Mega Test, the Ultra Test, the Power Test, and the Titan Test in identifying highly intelligent persons – despite being compromised?

May: There is evidence that uncompromised tests work better.

Rosner: Maybe, some smart people still trickle in. The Mega Test has been compromised since, probably, the late '90s or the internet made it possible to contaminate the questions by throwing around answers in chat rooms.

The Mega Test was the most successful in finding high-IQ people because the most people took it when it was published in Omni magazine. 4,000 people took it. It's more than any other test ever.

Which means, though, more people have taken the Hoeflin tests than tests by any other author, though probably a strong second and possibly somebody who has overtaken Hoeflin because he has written dozens of tests is Paul Cooijmans, who has been writing tests for decades and has cranked out quite a few.

Some of his tests have certainly been taken by more than 100 people. In the aggregate, thousands of people must have taken Cooijmans tests. With the success of the Hoeflin tests, they have found, depending on the cutoff, hundreds of high-IQ people.

Some of those people got together and some people were mentored by other high-IQ people, and had their lives improved, including myself. So, the success of the Hoeflin tests is the large numbers of people who have taken them.

For years, I, and sometimes with partners or being asked to consult, pitched TV involving high-IQ-type competitions. The same kind of shit as Project Runway or American Idol. A talent search, but instead of for fashion designing or culinary skill or singing skill, it was for raw intelligence.

This is an idea that comes to people not infrequently, but just has never been turned into a show. But if you had a show that did that, that would be the most successful project ever to find high-IQ people because millions of people would see the show and tens of thousands of people, if there were high-IQ tests associated with the show, would try those tests.

But that project has never happened, which I think is stupid because reality shows are about following assholes around with cameras and there are plenty of high-IQ assholes. Not as a percentage of high-IQ people who are, as I said, mostly decent, normal-ish people.

But if out of 100 people who have managed to score 160 on an IQ test, there are probably a half-dozen who you could productively, entertainingly follow around with cameras.

Cole: First of all Ron Hoeflin is a talented question framer. Next he spent a lot of effort validating his questions. Finally he normed them several different ways.

Jacobsen: In principle, what is realistically needed to test between – let's say – 4 and 5 sigma above the norm, reliably and validly?

May: Perhaps advanced AI can be used to develop significantly improved high-range intelligence tests. Other neurobiological methods of assessment of the general factor of intelligence, 'g', may eventually make IQ tests obsolete. For example, measures of biological traits such as pitch discrimination ability (of sound frequencies), among other such physical measures, have been found to have surprisingly high correlations with general intelligence. This may be the way of cognitive ability assessment in the future.

Rosner: You need experienced test-builders. You need a decent amount of people to norm the problems on, to make sure the problems can actually measure high-IQs. You need their other scores to see what scores getting those problems right correspond to.

As I said, you need some kind of widespread exposure. You have to let hundreds of thousands of people know that the test exists. Ideally, that it's something fun and/or cool to do.

Another condition is that it would be really, really helpful if the test took less than 20 hours to take. It would be helpful if someonecould spend 20 hours or 10 hours on the test and score near the ceiling, which is not a common thing among these tests.

Cole: To avoid spoilage you need question schemas, not single questions. Then you need a way to automatically collect many samples. Presumably this would be on the Internet. A group of Mega members is working on this. Contact me if you'd like to help [Ed. chris@questrel.com.].

Jacobsen: What is the principal design of the Adaptive Test, inasmuch can be stated at this time? (Is this series the first announcement of the test, by the way?)

Cole: Cf www.mental-testing.com. There are some articles in Noesis. Let me check with the team.

Jacobsen: What other extraordinary high-I.Q. societies have been observed by you – the highest, most inclusive, most exclusive, the most multi-planetary, least reliant on D.N.A. prejudice, most non-carbon-based, und so weiter?

May: The Plurality IQ Society

Top 0.00000000000000000000000 … % of Multiverse

Previously the highest-IQ group founded was the Aleph Society, which sought to have at most fewer than one member per Multiverse potentially qualifiable. However, the Aleph is found to be insufficiently selective in its admissions criteria for several reasons. First, it only considered 3 dimensions of space and 1 dimension of time per universe. We feel that it is necessary to include all theoretically possible multiple dimensions of spaces and of times per universe of the Multiverse. (For multiple-time dimensions see, e.g.: https://en.wikipedia.org/wiki/Multiple_time_dimensions , https://arxiv.org/abs/0812.389 , https://bigthink.com/surprising-science/there-are-in-fact-2-dimensions-of-time-one-theoretical-ph ysicist-states/)

Secondly, the Aleph only sought the highest IQ 'individual', including AIs, in the Multiverse 'now', i.e., at only one point in 'time' relative to one (1) observer, the Wormhole Officer (formerly called the Membership Officer). To remedy this we 'now' recognize that to whatever extent possible technologically, the Wormhole Officer must be a time traveler.

Thirdly, it is not sufficient that our psychometric instruments selecting at the Aleph level be culture free. Our IQ tests must also be genome free, i.e., free of any genetic influences upon performance. Speciesism is even more common than racism and gender-bias. We seek genetic justice in our member selection testing criteria. For example, in the past and even today, species with brains are unfairly advantaged over species without brains, including, of course, AIs. Why should an Isaac Newton have an IQ advantage over a slug, simply because a Newton has a brain? This obvious bias must be eliminated.

NB: All of the non-members of the Plurality IQ Society are Full Non-members and Official Non-members.

Jacobsen: What is the system of thought or the psychometric philosophy behind associative horizon?

Rosner: In my mind, when you get hit with a hard problem, one that might take more than ten hours to figure out. Part of it is how many different angles can you come up with on the problem. How many parts of life can you apply? How many possible analogies can you apply? How many keys are on your key ring to approach the problem?

When he talks about associative horizon, it is how many associations can you possibly come up with, with the symbols or whatever, that constitute the problem. To some extent, taking one of these high-range tests is profiling the author, trying to figure out, maybe, them, Hoeflin problems have a Hoeflin flavour to them, let you know if you are on the right track. Other test makers have flavours similar to them too.

It may be similar to their culture, say. The person building the problem found something in their world and boiled it down to an analogy. There is a popularish puzzle that is 7 d in a w." You have to figure out what the "d" and the "w" are. It's 'days in a week.' The problems can get tough. Another easy one. "5,280 f in an m," 'feet in a mile.'

So, "106 billion p who e l." The "e" "l" is tough. You have to figure out. It is 'people who ever lived.' So, for a lot of IQ problems, they have at least some aspect of that. Decoding, figuring out what the symbols represent. Then it is an exercise in figuring out what could the "p" and the 'p in e l' stand for.

"6*10^23 As in an M." My numbers might not be right. But 'atoms in a mole,' it is a test of cultural literacy. Often, there is further manipulation done to the symbols, so you have to work through two or three transformation or link two or three transformations to figure out the problem. It is how much cultural literacy do you have or do you give yourself, and then the flexibility for combining these things.

It is how much different stuff can you bring to bear on a fairly obscure or convoluted problem.

Jacobsen: How did you first come to find the Mega Test?

May: Actually I don't remember. It was about 40 years ago. I probably met Ron Hoeflin through my membership in the Triple Nine Society. This was probably my initial connection to the Mega Test.

Rosner: Some guys in my dorm told me about the Mega. I must've already been IQ braggy. Yuck.

Cole: Saw it in Omni Magazine.

Jacobsen: What were the claims about the Mega Test – and your score(s) in each section on it – by Ronald Hoeflin, the media, and others?

May: Ron Hoeflin told me that I was the 2nd person to obtain a perfect score on the 24 verbal analogies, I believe. I think Marilyn Vos Savant was the first. I certainly didn't tell many people, beyond my girl friend. I remember showing a copy of the Mega Test to one young woman, thinking she might be interested. She just laughed and laughed. Neil Blincom of Mr. Pecker's original, illustrious National Enquirer tried to interview me once when I was Membership Officer of the Triple Nine Society. I pondered this offer deeply for a fraction of a second. I remembered Chris. (never forget the decimal point) Harding's interview, "World's Highest IQ Genius is an Unemployed Janitor" and decided not to be interviewed. I avoided the media.

Rosner: So, the claims were the Mega was the world's hardest IQ test. By hardest, having the highest ceiling, the score a perfect score would get you, for instance. I think after the sixth norming, after Ron looked at 4,000 test submissions that came through Omni. I think the ceiling became 190 S.D. 16 or a little over 5.6 sigma. The first time I took it, I got a 44, which was 23 verbal problems right and 1 wrong and 21 math right and 3 wrong. I took it a second time and got a 47, which was 1 math wrong, I think. It doesn't matter whether math or verbal; I got 1 wrong the second time.

What does that translate into for me, after the fourth or fifth norming, my 44 wasn't high enough to get me into Mega. Marilyn herself turned me down for admission. My score might have corresponded to 172. Then after the sixth norming, after all these scores came in, I think a 44 got you a 180. I think the Mega cutoff is a 176. There you go. The 1-in-a-million level. Next question.

Cole: Omni called it the "world's hardest IQ test." Interpretation of scores can be found in Hoeflin's normings.

Jacobsen: How does the internet complicate legitimate testing in the high-range?

May: The internet facilitates cheating on tests and meeting other cheaters to work with.

Rosner: The Mega came out in '85. The Titan, the sequel to the Mega, came out in '90. Most people got on the internet in the mid-to-late-'90s. For those tests, it complicated and contaminated them because people went on message boards and threw answers around. Some of which were correct. That was problem one. Problem two was once Google came along; you could put in the words to the analogy and the fourth word would pop up. The analogies were half of the Titan and the Mega.

The 24 verbal problems were all analogies of the type "find the fourth word." Most of those could be instantly solved using a decent search engine. Tests are different. The Cooijmans tests, which I consider the most challenging of the internet era tests can't simply be solved by plugging things into a search engine. You still have to figure a lot of shit out. The most general issue with these tests and the internet is just sharing answers. Beyond that, it is a pain in the ass to make sure that the problems on the test can't be solved through easy searches.

Chris (Cole) and his group of people, who are working on this test that are resistant to having answers shared, are working on tests that give each test-taker the same general problem, but the specifics of the problem are fresh. So, somebody else's answer on this problem is not going to help you because, even though the problem should score the same – getting it right should reflect the same IQ level, you can't just post what you got on answer 12. They've been working on that for well over a decade.

It's coming along. Anyway, next question.

Cole: The Mega and Titan tests have been spoiled on the Web. The Power and Ultra tests are at risk.

Jacobsen: Some, in fact more than a few, claim extrapolations well beyond the norms of the mainstream tests, e.g., the WAIS and the SB, which cap out at or around 4-sigma. Assuming legitimacy of the claims, then, the individuals would be highly intelligent, but the claims can range between a little over 4-sigma to 6-sigma. How is this extrapolation generally seen within the high-I.Q. communities at the higher ranges?

May: I don't know how other others generally perceive unsound or bogus extrapolations of IQ scores.

Rosner: I think the skepticism of super-high scores is generally more for specific claims than for the entire idea of being able to have an IQ that high. I think most people in the high-IQ community believe it is possible to have an IQ close to 200. But I think most people also have a reasonable idea of the rarity of scores like that. Adult IQs, the deviation scores, are based on a bell curve, where between 0 and 1 standard deviation, you have 34% of the population in a bell-shaped distribution for something like height. Between 1 and 2 SDs, you've got 14% of the population. Between 2 and 3, you've got about 1.5% of the population. Between 3 and 4, you've got roughly one-half percent of the population.

Let's see, about 4 SDs, that's only one person in 30,000 should score above 4 SDs. One person in 3,000,000 above 5 SDs. What is it? 1 person in 750,000,000 above 6 SD or so; somewhere, I've fucked it up, according to the standard bell curve. People also like to say that at the very far ends; there are more outliers than on the normal bell curve. That there are more high-IQs than would be given if it were a perfectly bell-shaped distribution.

But even so, you shouldn't see more than a half-dozen or ten or twelve or whatever, people, with scores above 6 SDs. So, Paul Cooijmans has the Giga Society, which has 7 or 8 members. It is for people with IQs that are supposed to be one in a billion. So, there are 8 billion people on Earth, 8 members of the Giga Society, so that makes a certain sense, but not really. That's as if everybody who could score at that level has taken one of his tests. That's just obviously not true. So, way too many people scoring at the one in a billion level. It's not like the Giga Society has 300 members.

Cooijmans is pretty rigorous in his norming and testing. So, if you have taken a Cooijmans test and scored at or close to the Giga Society, legitimately, Cooijmans has written in the past about people's attempts to cheat on his tests, but I don't think there has been a successful attempt in decades. So, people are pretty accepting that if you get a Giga level score on his tests; that you're legitimately pretty smart. The claims of super high-IQs, there are legit claims based on performing well on ultra-high IQ tests or kicking ass as a kid on a test like the Stanford-Binet or the Wechsler. Someone can say, "As a kid, I scored a 200," or something.

That's another thing I won't go into. People who claim high-IQ scores and are lying are generally not sophisticatedly lying. They're saying something that cannot hold up at all. I don't know if there are many or any sophisticated lies about having a super-high-IQ. So, then there are people outside the high-IQ community who are skeptical about the whole thing, but no one is really worried a lot about it, because: who gives a shit?

Also, if you want to say something, or know something that I'm not aware of, that contradicts what I'm saying, go ahead.

Cole: Hoeflin's norms all involve some extrapolation. I find it reasonable up to the mega level (about 4.75 standard deviations).

Jacobsen: Following from the previous question, what seems like the motivation behind making claims well beyond the norms of the most used mainstream I.Q. tests?

May: It's a shame Einstein did physics. He could have been on Facebook (now called Meta, I guess).

Rosner: Going off my own experience, I kind of felt like a loser based on when I was about 20. I'd fucked up a lot of opportunities for myself. Then somebody told me about the previous world's hardest IQ test, which was a Kevin Langdon test. It ran in Omni or Games Magazine. I took it and scored 170. I went, 'Wow, that's a good score.' When Mega came along, I took that. I liked that validation that it gave me. Even though, it is a ridiculous thing. I kind of feel like it might be analogous to a guy who can bench press 500 lbs.

It's kind of a goofy thing. You wouldn't tell that guy it is goofy to his face, but the Sven Magnason. He is 6'4" and weighs 310 lbs. and eats 200 grams of protein a day to get that or support that huge bench press and has hypertension and his joints will be fucked in 10 years. It's a kind of a goofy thing. It is amazing the guy can bench 500 lbs. It is this ridiculous thing. It is a very obscure sport. Sven Magnason is not playing in the NFL for 1.8 million USD a year. He probably works in a warehouse and does strength training on the side.

It doesn't translate into the kind of fame or success that you might want. So, it is a niche kind of sport.

Cole: Vanity is one motivation.

Jacobsen: What are some of the more egregious I.Q. claims in 20th century by groups and by individuals? This is a free forum.

May: In the 20th century — maybe being the smartest man in America was a fairly egregious claim. Top 1 per billion high-IQ societies may qualify if such came into existence in the 20th century.

Rosner: I don't know. Anybody can go on the internet and type whatever they want. One of the craziest claims I saw I mentioned before. Somebody had a site or has a site claiming Jesus had an IQ of 300. The idea that somebody with the deep wisdom of Jesus meant Jesus had a huge IQ. His estimate based on nothing: If smartest people have an IQ of 200, then Jesus must have an IQ of 300. William Sidis, people claim 259 based on extreme achievements as a young person, at least it is based on his history and is a fairly earnest attempt to estimate a very smart young man's IQ.

It is kind of egregious and not based on him being tested. Oh! Some of the most egregious are in the last 15 years; some insane moms, one mom out of Colorado, maybe 18 years ago, got a hold of the answer key to an earlier edition of the Stanford-Binet. Stanford-Binet gets revised every 15 or 20 years. I don't know. You can still find psychologists who will give an earlier version. In the stacks of libraries. Probably, the Norlin Library at the University of Colorado, she found an earlier editions, found an answer key. Then taught her kid all the answers, so, that kid scored, at age 3 or 4, like a 10-year-old, which, the way they calculate childhood IQs, gave him an IQ well over 300. She tried to get herself and her kid famous off this.

It, eventually, fell apart because the kid did not have a 300 IQ. So, that is pretty egregious. But! Doable if you're not an idiot about it, I believe. But anybody who would do it would be a kind of idiot. First of all, I don't know. How much would a 4-year-old be into it? But if you took a 6-year-old and got a 6-year-old into it, "We're going to ride this pony into a T.V. show, your acting career." It has never happened, but it is not impossible. Because Alicia Witt was a child actor, an

actor now. Great actor and great kid actor, one of the things that makes for a great kid actor is a 4-year-old who can read.

Because if you can give a 4-year-old – Alicia Witt could read at 3 – a script and the kid can read the script and memorize the script rather than having to be told shit line by line, and if the kid is smart enough to do that, then the kid is smart enough to take direction. Alicia Witt was at least a kid actor because she was super fucking smart. So, I'm thinking if you had a motivated 6-year-old and a creepy parent. I even started working on a screenplay on this or thought about it 30 years ago as a good plot. Like a lot of shit I do, I didn't do anything with it, except the mom did it and a shitty job in real life.

The right combination of psychopathic parent and bright, motivated kid. That team could believably sustain the bullshit that that kid has an IQ of 300+ for quite a while. Although, nobody has done that. Yes, that would be egregious.

Cole: Before they were banned by Wikipedia, there were many articles by groups making incredible IQ claims.

Jacobsen: What seem like the big lessons in debunking phony I.Q. claims from the 20th century?

May: "The first principle is that you must not fool yourself and you are the easiest person to fool." — Richard P. Feynman

Rosner: [Laughing] A lot of stuff underlying a lot about high-IQ is "Why?" Why claim to have a high-IQ? Why work your ass off to get a super high score on these tests? Why sweat debunking it? In retrospect, you can see why you might want to hold people who might claim super-high-IQs up to scrutiny, at least given Raniere. The NXVIM sex cult, swindler of the Bronfman's who is in prison for life now. One of the pillars of his duping people was using a high score on the Mega Test to claim to be one of the smartest people on Earth, though he didn't really push it.

Because once he gathered enough acolytes, I don't know enough about him to know how often he dragged out his IQ. But it seems that once he was surrounded by dozens of followers; that he didn't need to do that. He could rely on his charisma and manipulation skills, and also being at the top of a pyramid of people with good manipulation skills. He was smart enough to recruit charismatic actors, TV stars. A couple actors from Smallville. People with actual show biz careers. One of his selling points and one of the selling points of Scientology can help you succeed professionally in shit where what it takes to succeed, like acting, can seem nebulous.

So, he didn't need to haul out his IQ a lot because he was surrounded by TV stars who were helping him recruit other people into his cult. He, certainly, deserved a lot of scrutiny, perhaps a lot sooner than he got the scrutiny. There's another guy who is pretty culty who has a bunch of acolytes who espoused a bunch of scary shit. So, that's one reason to scrutinize claims of super-high-IQ because people can be up to no good, but those people are fairly rare. Of the 60, 80, 100, people who have qualified for the Mega Society over the past 40 years, 95 or more percent of them are completely normal, undangerous people.

The biggest danger might be that they might be really funny, like Richard May, is a completely decent guy who happens to be extra smart and extra funny. Super-high-IQ people mostly aren't to be feared. What were we talking about? I always talk myself way away from the question. [Ed. Question repeated.] That, I guess, let the babies have their bottles for the most part, let high-IQ people be high-IQ people, it doesn't hurt anyone, except for a few cases. Those involved in IQ fraud, the fraud is pretty transparent.

Most of the high-IQ lying is some desperate asshole who is 25 and going to undergraduate parties at his school. That guy finds a freshman girl and says, "Oh, people don't understand me. I have a 205 IQ. I graduated high school at age 5." It's that abject bullshit. There are more sophisticated attempts, but not that much more. Because the payoffs are pretty low. Even lower than getting a hand job from a freshman girl, the end.

Cole: "It's hard to be right." — Richard Feynman

Jacobsen: What would you define as fraudulent activity in a high-IQ community or an individual?

Rosner: Making claims that you know aren't supported by your performance on tests.

Cole: Fraud takes many forms just as it does in common law. Because of the Internet, tests with fixed questions are particularly vulnerable to cheating.

May: I have nothing to add.

Jacobsen: What would you define as messianic posing in a similar regard?

Rosner: If you end up with a cult, that's messianic posing.

Cole: The common language definition of messianic behavior will serve.

May: I have nothing to add.

Jacobsen: Similarly, what about criminal behaviour?

Rosner: If you end up in jail for the rest of your life, if the FBI has a thick dossier on you because you are considered a potential threat in certain ways, that's criminal behaviour. The FBI has dossiers on lots of people because, historically, the FBI has done good things and asshole things.

So, if they have a dossier on you, because you're a legitimate psycho who has the potential to do bodily harm to people for some weird political reason, then there you go.

Cole: Again I have nothing to add here to the common language definition of criminal behavior.

May: I have nothing to add.

Jacobsen: On the Mega Test, why was the three interpenetrating cubes problem seen as the most difficult?

Rosner: It is widely agreed that the three interpenetrating cubes problem was the hardest problem on the test. So, the problem that is agreed upon as likely being the correct answer has not, as far as I know, been proven to be the correct answer.

Interestingly, you can look it up. It depends on what shit is online. But at various times since the ‘90s, it has been agreed upon that the correct answer is floating out there. But you can’t be sure that you’ve found the consensus correct answer.

But the figure, the geometric figure, that corresponds to the consensus correct answer can be found in popular culture, but I won’t tell you where.

Cole: It’s the only problem on the test where the answer that Ron accepts has not been proven. There are a few of these on the Titan.

May: It was the certainly most difficult, but my spatial ability is not sufficiently high to understand why this is so.

Jacobsen: Above 4 standard deviations above the norm, why should there be more scrutiny more than any other cutoff?

Rosner: Isn’t there some claim that “extraordinary claims require extraordinary evidence”? You could argue that because claiming to have one of the world’s highest IQs gets you more than claiming to have a 120 IQ.

In practical terms, not so often, it can get you on a quiz show. It can get you on the cover of Esquire magazine. It can get you interviewed. It can get you on TV. It kind of got me laid once. I was going to get laid anyway. But it was part of that package that got me laid, I guess.

Cole: A credible high range score requires credible high range test questions, which are hard to formulate and norm.

May: I have nothing to add.

Jacobsen: What was the hardest IQ test you’ve ever taken in the high-range? What lesson can be learned for test-makers from this?

Rosner: I say that I’ve had a lot of success, but I’d say that I’ve had the most difficulty with Cooijmans’ tests. Because he brings in stuff from a lot of areas. I don’t want to say too much about his tests because he doesn’t want people talking about his tests and helping other people.

But by the time the Mega Test had been published in Omni, it had been through a number of revisions with hinky problems getting knocked out or revised until they were clear and bullet-proof. The answers were tight. I think Cooijmans talks about the pleasure of when an answer clicks into place. That click of satisfaction of when you know you found the answer.

I would say that on some of Cooijmans’ problems. The click is, maybe, not as loud as on some Hoeflin problems. On Cooijmans’ problems, you can find some really good answers that aren’t as good as the intended answer. That’s, maybe, the mark of one type of really good ultra-high-IQ test.

That there are stopping points. On multiple choice tests, those are called distractors. There are answers among the choices that seem right for various reasons if you’re taking desperate stabs at an answer.

On high-IQ tests, you can come up with answers that make a lot of sense. But do they make as much sense as the intended answer? No. But you've fallen for an inferior answer. On tough tests, a lot of problems on hard tests are finding the signal among the noise.

I'm writing a book in which somebody or the recipient of what he thinks is a coded message, thinks that it is a true message because it is based on the first letters of four consecutive sentences. That spell out a word.

The odds that this would happen by chance are 26 to the 6th power, which is 676 squared, which is 400,000 to 1. Then you have to knock that down because there are a zillion four-letter words. So, anyway, the odds are tens of thousands to one that it's not a coded message, especially since it is specific to the character situation.

So, the character reasons that it is likely a true signal. And on a tough IQ problem, you'd like the numerical coincidences to have an unlikelihood of, at least, 1 in a 1,000. When you look at a number sequence, you see a pattern. Then you say, "What are the odds that this pattern would arise by chance?"

On some super-hard IQ problems, there are more than one pattern to be found. Again, you have to ask yourself, "Was this intentional or accidental?" A tough-ass IQ problem really pushes the limit in finding the signal among the noise.

Cole: The only high range test I took was the Mega.

May: The Mega Test and the L.A.I.T. are the only high range tests I've ever taken.
I did not distinguish myself on the latter.

Jacobsen: Is IQ declining in importance now?

Rosner: IQ as IQ is declining in importance because it is a product of the middle of the 20th century when people really believed in it and used it to skip kids a grade, or not, to put them in gifted classes, get admission to magnet schools.

At some point, probably in the '50s, you might be able to get laid by your IQ. Since debunked, it has a greasy feeling about it, weirdo, creepazoid. The Cal. State schools, today, decided to get rid of the ACT and SAT altogether and the SAT is an IQ surrogate.

They decided it is not helpful, not worth the shit people go through to prepare for the tests. We can see enough about a student without some IQ surrogate in their admission packet. I'd say intelligence is increasing in importance because we are tiptoeing up to artificial intelligence.

That when we talk about AI – and AI is a misnomer right now; AI means "machine learning." Eventually, AI will mean "Artificial Intelligence." We will need ways to mathematicize and to come up with metrics of the power of thought in brains and in other stuff.

So, old school IQ declining; new school AI shit increasing.

Cole: IQ seems to be about as important now as it was when I was young. The SAT has some problems because it has become easy to improve a score via tutoring, but that is being addressed.

May: There is a theoretical possibility that Nature, specifically natural selection might not be entirely "politically correct." Theoretically there could be differences among human groups that evolved under different conditions. E.g., If only females could bear children, then males would be the expendable 'gender'. A small number of healthy males could impregnate a large number of females and the group would survive. A large number of males, if males did not bear children, and a small number of females would not allow the group to survive. Hence, there could be more variability among males, including cognitive variability, because males would be more expendable, than among females, i.e., there would be more male 'geniuses' and more male idiots.

Fortunately we now realize that there are no biological differences between males and females. Gender is a purely social construct. We now realize that men can menstruate and have babies too, if given a chance. The only important differences are among large numbers of pronouns, all referring to identical nouns.

Chris Cole, Trip Report

2024-08-15

Original publications here, October-November 1986.

High IQ societies usually attract oddballs and cranks, so I have never joined one. However, the Mega Test was so difficult that I figured it would weed out these people, since they wouldn't have the patience to work out the answers. This is also why they never make important contributions. However, all of this was pretty much conjecture on my part, so it was with some trepidation that I set out to meet some fellow members of the Society. I figured I would either meet the crème de la cranks, or a bunch of people more or less like myself.

During September, I met with four fellow members of the Society: Jeff Ward and Dean Inada in Southern California and Ron Hoeflin and Ray Wise in New York. To my great relief, I found that they were not cranks. Not one crank idea was proposed during any of the several hours of discussions. The ideas that were discussed were fairly examined from all sides, and people were willing to change their opinion when presented with sufficiently strong evidence. It was very comforting.

We discussed the solutions to Trial Test A and formulated a consolidated solution set (we could not solve problems 33 through 35—these are still unsolved as of this writing). We agreed that it is important to expand the Society and that tests such as the Mega Test are the appropriate vehicle to do so. We had several suggestions on how to minimize cheating on the tests:

1. Don’t publish the test.
2. Have the person requesting the test sign a contract stating that he or she will not reveal the contents of the test.
3. Change the test every year.
4. Specify that admission to the Society will require an interview that will involve some follow-up questions, even though this may not be true.
5. Tell the person before he or she requests the test that the test will require a considerable amount of time, and then be strict in requiring that the test be returned within the time limit (say, three months).

We also discussed several projects for the Society. Jeff suggested a forum (television show? magazine?) for critical, objective assessment of arguments on both sides of issues of public interest. I suggested a long-term project in the area of cellular automata and artificial intelligence. We all agreed that there would be no shortage of ideas on projects, nor any shortage of energy and talent to apply to the projects. All in all, they were two very enjoyable meetings, and I look forward to more.

Rick Rosner, On Stupidity

2024-08-15

Original publications here, December, 1994.

While we debate the stupidity of material in Noesis, lemme say a few things about dumbness in the outside world. Observation one: Fifty years ago, the Era of National Stupidity reached its peak. We are now in the Era of Individual Stupidity. During WWII, nations were psychetic, but the individuals that comprised those nations generally behaved themselves according to the rules established by their crazed leaders. Today, the world's largest nations generally behave with some restraint, but the individuals in those nations misbehave.

I blame an increasing population and productivity for widespread dumb behavior. Since WWII, the U.S. population has doubled, and productivity has increased five or ten times. This is too much productivity. There's not enough stuff to do, and people must fritter away their time, going to college, watching cable, playing video games, filing lawsuits, pursuing meaningless (and usually vicarious) sex. This is fine with me, except that, as a professional moron, I can't keep up with all the amateur manna.

Self-destruction through individual misbehavior is certainly preferable to the destruction of populations through national aggression. It's fun to wonder when this trend will lead. (Incidental) Observation two: In political propaganda, 'Where it will lead' is the type of argument most frequently made. Most court cases, most political decisions, are pithily and can be seen as significant only through the magnifying glass of trend-mongering—'If stuff like this keeps happening,' the argument goes, 'we'll end up in some politically-extreme dictatorial dystopia.' (That's how I feel we're trending now under the Republicans, but I should know better.) Most trends exist only to fill newscasts. Piddliness in one direction is usually scuffed out by a succession of other oddly trends.

But, maybe individual media-abetted techno-sexual-criminal foolishness is an actual trend. Then things can only get more interesting. With more people with more resources to create their own little worlds, each individual slice of life, each biography, is going to be thinner, more tweaked, a more distant random divergence from some 1950's average. And, sociobiologists et al like to argue that altruism is genetically based. They do the math and show how genes survive better under cooperation. Observation three: I bet there's some other math to be done showing that when a species is too successful, some genes survive better using chaotic, violent strategies. EvGybusly knows when too many rodents are crammed in a cage, they engage in antisocial behavior. There's gotta be some sociobiological math behind that.

Quick review of I.Q., in which Walter Matthau plays Albert Einstein—Much of the movie takes place at Princeton's Institute for Advanced Studies. Podolsky and Kurt Godel get lots of screen time as Einstein's sidekicks. So during the first few minutes, I was pretty excited. But the movie is real dumb, even for non-physics people. President Eisenhower comes to campus to congratulate Einstein and Tim Robbins for developing cold fusion. Einstein rigs a car to malfunction by remote control. There're enough moments of oksyness to keep you interested, but the movie ends with a messy cluster of coincidences and unlikely behavior. You might impress a date by pointing out all the wrong stuff, but you'll probably just sound annoying.

Seneka, Point of View

2024-08-22

Scott Douglas Jacobsen: Seneka, could you briefly introduce yourself and your areas of expertise?

Seneka: I think it's accurate to call myself a polymath. I have worked and have experience in dozens of topics such as business, hypnosis, magic, art, chess, research, writing and more.

Spanish is my native language and I self-learned English by reading some books or watching online videos. Just a heads-up: my English may sound a bit "basic".

Jacobsen: When did this interest in test construction truly come forward for you?

Seneka: It all started from a wider curiosity about the nature of intelligence and how we can understand it better.

In traditional tests, I achieved the highest scores, which made me question the limitations of these tools in measuring the complexity of human intelligence. This is where my interest in alternative evaluation methods began, allowing me to include my understanding of multiple intelligences in test design.

Jacobsen: What were the realizations about the tests earlier, and then the need to develop yours?

Seneka: Most IQ tests tend to focus specifically on certain cognitive abilities like reasoning, working memory, or crystallized knowledge. While these are definitely important and, in my opinion, accurate, they don't show the full range of human intelligence.

Jacobsen: What was the origin *and* inspiration for the creation of this test – the facts and the feelings?

Seneka: The origin of my test, Point of View, comes from both research and a more personal goal. In fact, POV is just one test in a collection that evaluates different high-level intelligences. On one hand, I aim to isolate and give more importance to divergent thinking in this test so it can be evaluated. On a personal level, I believe that we don't value the intelligence of many people because we only measure the logical or rational part.

Jacobsen: What do you mean by "evaluating divergent thinking"?

Seneka: I think Howard Gardner was very right with his theory of multiple intelligences. In fact, my idea of intelligence is similar, but with a classification of intelligences that is not connected to educational, cultural, environmental, and social factors.

Divergent intelligence is what allows a person to look at a problem from different angles. For example, an "X" might just be the letter X for someone, but for someone with high divergent intelligence, the same X could mean a mistake, a number, two lines, a cross, a destination on a map, a selected option, a chromosome, an adult content rating, a prohibition, and many more things. The POV test specifically evaluates a person's ability to see concepts through divergent and lateral thinking. Once you find the right perspective, the logic to apply is really simple.

Jacobsen: What skills and considerations, in an overview, seem important for both the construction of test questions and making an effective schema for them?

Seneka: Creating effective questions requires the use of my own divergent intelligence. It's important to design items that can differentiate between different levels of ability. To do this, I need to analyze the most common uses of each element and progressively move away from the more unfrequent perspectives.

Jacobsen: With Point of View, why focus on a matrix design?

Seneka: I like matrix tests because the information is contained within the problem itself. I also know that people feel more motivated to take a test when it looks interesting. The structure is very good for "misleading" the test-taker. It hides the perspective from which to approach the problem behind hundreds of possible logics that lead nowhere. With the matrix design, I can better evaluate a person's ability to think laterally and remove the noise.

Jacobsen: What do you mean by a "high-range matrices test"?

Seneka: These are tests aimed at people who probably perform at a very high level on traditional intelligence measurements. In my test, I try to isolate and give more weight to divergent thinking over logical or rational thinking. So, it's possible that there may be some differences between IQ (from the traditional test) and DQ (divergent quotient).

Jacobsen: As the aim is to measure divergent and lateral thinking, how does this style of matrix design differ from more traditional mainstream tests like the Raven's Advanced Progressive Matrices?

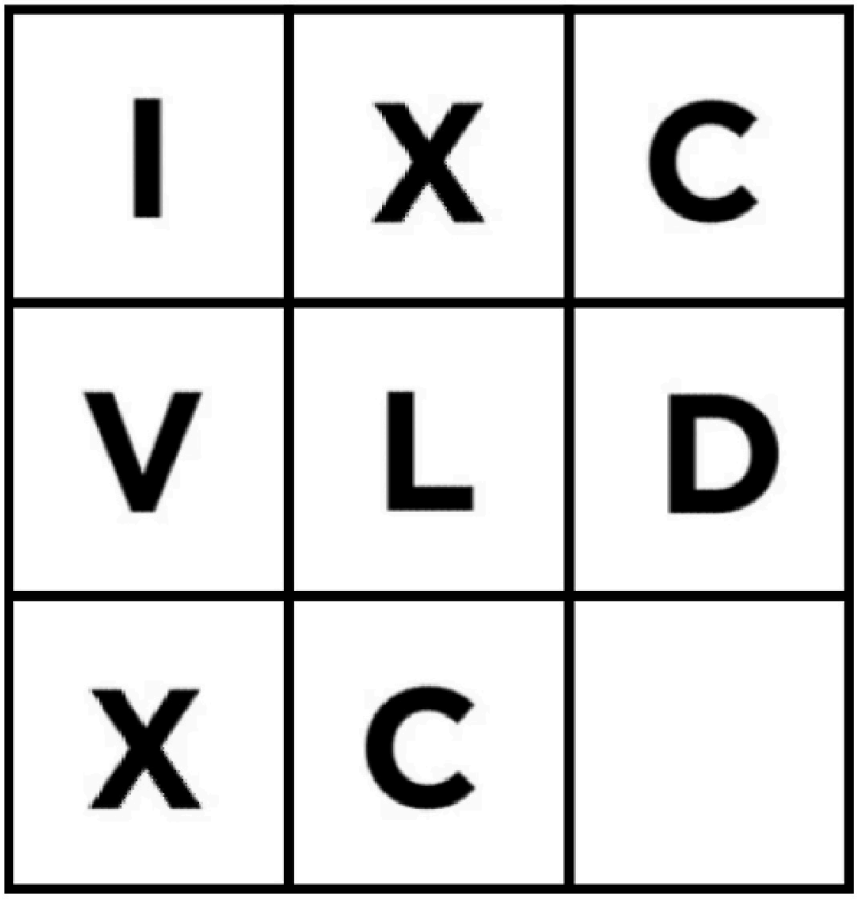
Seneka: While tests like Raven's Advanced Progressive Matrices focus mainly on logical-deductive reasoning, spatial vision, and pattern recognition, POV introduces elements that require lateral thinking. The design of my test intentionally includes ambiguity to push the test-taker to explore other approaches when the "logical" approach doesn't lead to any solution.

Jacobsen: When trying to develop questions capable of tapping a deeper reservoir of ability, what is important for spatial and matrix test type of questions?

Seneka: Even though other intelligences play a role in a high-level test, the goal is to isolate divergent intelligence until it becomes the most important factor for solving the problem. In fact, the test-taker should indicate how they reached the conclusion for each item. If they managed to find the perspective from which to solve the problem, it validates their divergent intelligence, even if the answer is wrong or they made a logical mistake.

Jacobsen: Potentially, what are roadblocks test-takers tend to make in terms of thought processes and assumptions around time commitments on these tests? So, they get artificially low scores on high-range tests.

Seneka: The biggest obstacle for participants is getting stuck on the wrong perspective. I'll give an example with an item that I developed and finally didn't include in the final test:



Here comes a spoiler. If you want to solve it, don't listen to this. Many people spend minutes or even hours trying to find out what logic is hidden behind these letters: they look at their positions in the alphabet, do math operations, or search for patterns. You won't find anything online if you search for those letters.

Only divergent intelligence allows them to "see" that the letters might not be letters and could be numbers. Once you think that, a simple search is enough to solve it. Maybe you even already know what numbers they are, and then you'll reach the conclusion even faster.

Jacobsen: What are the difficulties in preventing cheating on tests in the online era?

Seneka: I had the task of creating original items that are not on the internet, which is quite complex, by the way. I also tested the exam with different artificial intelligences, and out of all the items, only one reached a correct conclusion. Some AIs might score between 115 and 125 on traditional tests. That's already well above the human average. However, in lateral thinking, AIs have no chance; we humans have irreplaceable intelligences.

Jacobsen: What are the most appropriate means by which to norm a test when, in the high-range environment so far, the samples tend to be lower?

Seneka: Invite people who already belong to high IQ societies. They will be more inclined to take this test. Also, it's important to warn that scores as low as three or four out of twenty-three

are possible, even for gifted individuals. Using methods like Item Response Theory (IRT) can improve the accuracy of norms, even with smaller samples.

Jacobsen: Pragmatically speaking, for really good statistics, what is your ideal number of test-takers? You can't say, "8,500,000,000."

Seneka: Getting several hundred or a few hundred participants would be optimal. This number allows for relevant analysis and keeps the needed precision to identify performance differences. To give a number, reaching 500 participants would provide a solid base and statistical validity.

Jacobsen: What tests and test constructors have you considered good?

Seneka: There are few high-range test constructors, and I don't want to argue with anyone here. In general, we could say they create harder versions of existing tests. They make tests that can be solved if you have knowledge of group theory, geometry, and mathematics at a very high level.

Recently, I discovered some tests by Laurent Dubois, and I found them very interesting.

Jacobsen: What have you learned from making a test?

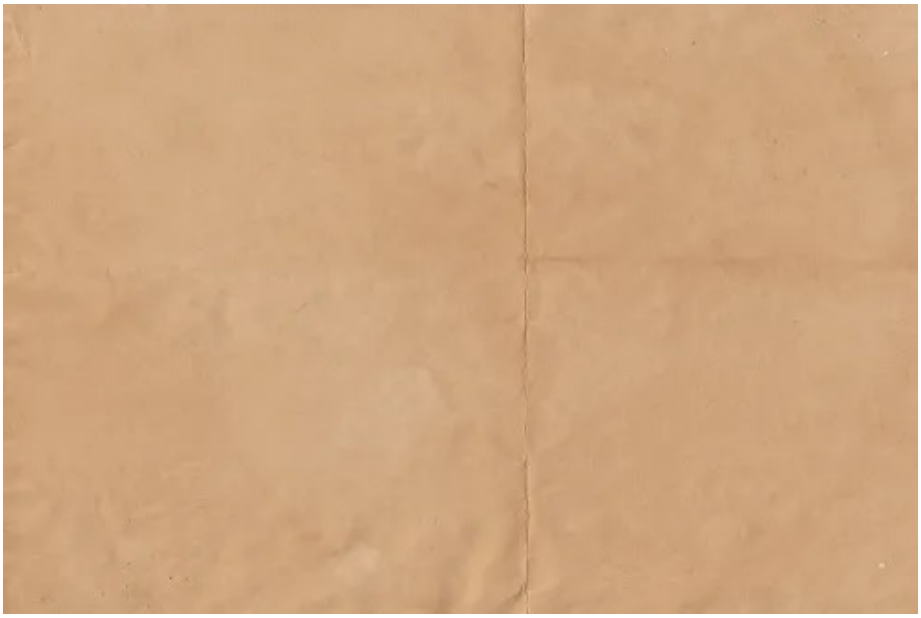
Seneka: It has been a very fun experience, more art than science. And it's helping me a lot to better define the concept of multiple intelligences, so I will continue doing this for other intelligences as well.

Jacobsen: Thank you for the opportunity and your time.

Seneka: Thank you, Scott. It's a pleasure to discuss these topics with you.

Chris Cole, Merger of Ultra and Short Form Tests

2024-08-22



*Original publication March, 1993 **here**.*

Ron Hoeflin has graciously consented to a merger of the Short Form Test and his work-in-progress, the Ultra Test. This means he has effectively donated the problems from his seven trial tests, which represents over a year of hard work. I propose that we call the merged test the Ultra Test.

Ron has convinced me to abandon the idea of a short-form test, in the sense of a small number of problems. There are two reasons for this: first, a small number of problems leads to statistical instability, and will make norming difficult, and second, by necessity, a short test would have all hard problems, which may be off-putting. In addition, a longer test will allow us to include several easy “aha!” problems, which will both entice and instruct the test taker. In other words, the easy problems indicate what kind of problems the hard ones are.

It is important for the test takers to understand that the problems are not amenable to exhaustive reference work or tedious calculation. Otherwise, they will abandon the test as too time-consuming. This explains, I think, the sharp drop off in takers between the Mega and Titan Tests. I think the audience of potential test takers was “burned out” by the Mega Test. With the Ultra Test, I hope to reinvigorate that audience as well as attract a whole new audience. There are many people who could qualify for the Mega Society if we could just get them to take the damn test!

In order to get a test published anywhere, it will have to be normed. This means it will have to be tried by a sample population. The only sample population readily available is the readership of Ron’s journals. Ron and I would like to publish the Ultra Test in the September issue of Ron’s journals. This will give us adequate time to collect and score answers by early next year.

Therefore, this is the deadline: all candidate problems for the Ultra Test must be received by September 1. So, please start thinking of "Ultra type" verbal and math problems and submit them.

Ron picked the 41 most discriminating verbal analogy problems from his trial tests. Ron calculates the percentage of high scorers who correctly answer a question and subtracts from this the percentage of low scorers who answer correctly. Thus, easy problems and hard problems have a low discrimination value. I further culled this list of 41 problems down to the following 12. The criteria I used are these:

1. Avoid reference exercises.

If the definition of the word is obvious from the analogy, but the word is obscure, the problem becomes a matter of searching reference material. This is not a test of intelligence; it is a test of who has the biggest thesaurus. I encourage all members to obtain a copy of Herbert M. Baus' *Master Crossword Puzzle Dictionary.* This book is the standard reference book of the National Puzzlers League and was able to answer 80% of the Quest Test. Barnes and Noble recently stocked up on these and sells them for $15. You can also order one from their 800 number.

2. Avoid idioms.

Idioms are not familiar to people for whom English is a second language. Native English speakers are a minority of the world's population. We should strive for a test that has a wider audience.

3. Avoid mythology and religion.

We should not expect Chinese speakers of English to know as much Western mythology as we know Chinese mythology. I know next to nothing about Chinese mythology. By the way, lest anyone think this is an overly harsh criterion, did you know that there are more students of English in China than there are speakers of English in the US?

4. Avoid wordplay.

A play on words is biased toward native English speakers.

5. Avoid quotations, titles, etc.

Again, these are culturally biased.

6. Avoid "A: synonym of A:: B:?" or "A: B:: synonym of A:?"

This is a catch-all criterion, meant to include analogies that do not fall into any of the above categories exactly, but which still are not so much analogies as they are definitions. The relation of synonymy is not a good basis for an analogy.

So here are the 12 new problems:

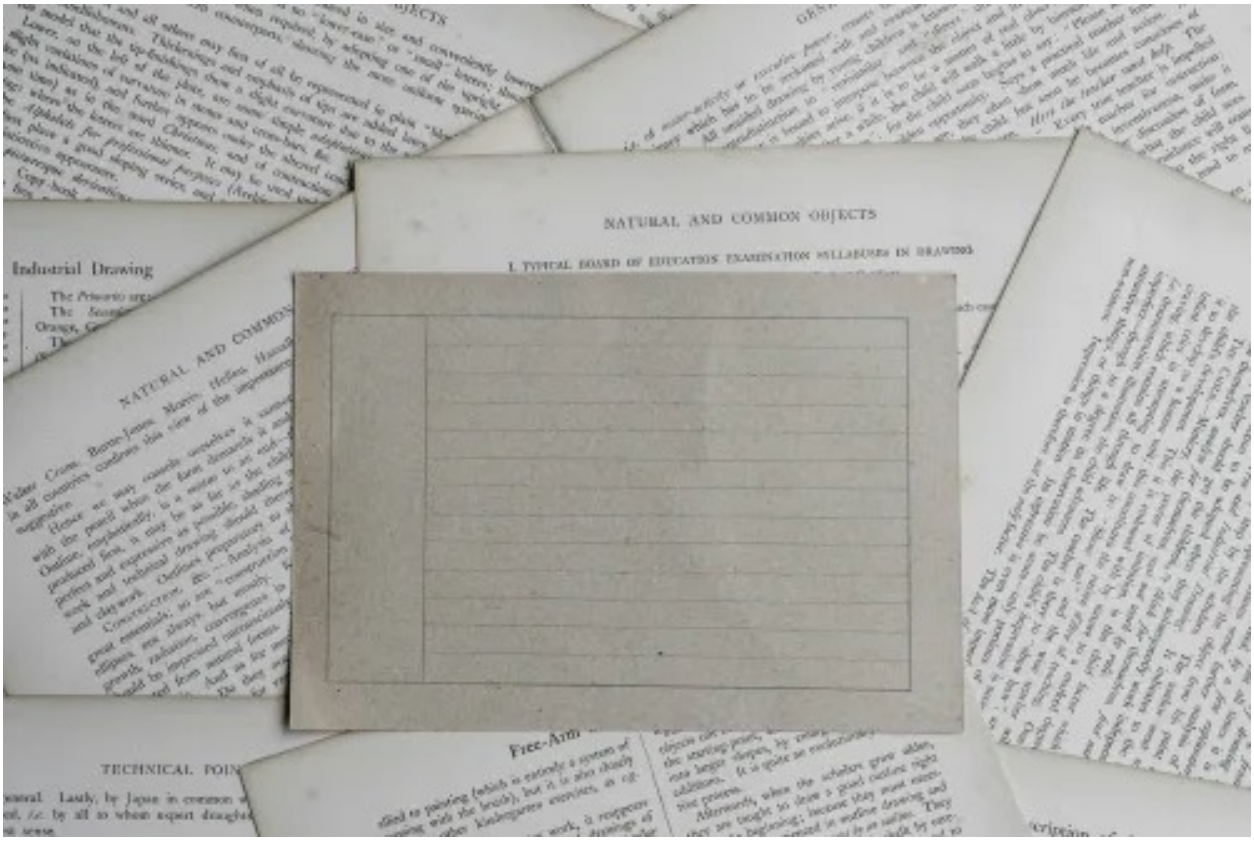
1. Space: Hyperspace:: Vector: ?
2. Image: Idea :: Hallucination: ?

3. Wind: Rain :: Typhoon: ?
4. Inward: Outward :: Infection: ?
5. Column: Row :: File: ?
6. Humbug: Bach :: Seek: ?
7. 38: Pyongyang :: 49: ?
8. Of ten: Factor :: Of magnitude: ?
9. Say: Hear :: Imply: ?
10. 2.54: Inch :: 3.26: ?
11. A, AB, B, BO, O: BO :: A, C, E, G, T: ?
12. Eggs: Grading :: Wounded: ?

In the next issue, we will present the spatial questions selected from Ron's tests, as well as all the other questions that will no doubt begin pouring in from the members who have been inspired by Ron's generosity.

Chris Cole, Why I’m Interested in Intelligence Testing

2024-08-22



Original publication April, 1993 here.

I have in hand a copy of the Terman Concept Mastery Test (Form T), which I got from Ron Hoeflin. This is the test used by Terman to track 1004 gifted children into mid-life. This test shows what is wrong with high-level intelligence tests.

Terman’s sample questions are fine:

- Shoe: Foot:: Glove: (a. Arm b. Elbow c. Hand)
- Kitten: Cat:: Calf: (a. Horse b. Cow c. Lion)

But it is all downhill from there. Here are some typical questions from the test, and what is wrong with them:

1. Proletarian: Worker:: Brahmin: (a. Bull b. Aristocrat c. India)
   This is a vocabulary question.
2. Bacchus: Revelry:: Ceres: (a. Agriculture b. Love c. Hunting)
   This is a mythology question.
3. Danube: Black Sea:: Euphrates: (a. Persian Gulf b. Red Sea c. Caspian Sea)
   This is a geography question.

4. Maoris: New Zealand:: Ainus: (a. China b. India c. Japan)
   This is an anthropology question.

Is a spelling bee an intelligence test? It may be the case that spelling is correlated with intelligence, but it is not the same as intelligence. I think Terman meant to produce a concept mastery test (which is a fine synonym for intelligence, as far as I'm concerned), but he could not think of enough hard analogies, so he got lazy and used hard questions from other disciplines.

The trouble is, we now can produce machines that can spell much better than we can (or do the simple information look-ups that are required by the above questions). But we cannot produce machines that can master concepts better than we can. So it is now interesting to quantify how good people (and ultimately machines) are at mastering concepts. This is why I am interested in true intelligence tests, and why I think the Ultra Test is worth working on.

I will give one example of a good "aha!" (or concept mastery) problem:

Start with a half cup of tea and a half cup of coffee. Take one tablespoon of the tea and mix it in with the coffee. Take one tablespoon of this mixture and mix it back in with the tea. Which of the two cups contains more of its original contents?

Answer on next page.

The two cups end up with the same volume of liquid they started with. The same amount of tea was moved to the coffee cup as coffee to the tea cup. Therefore, each cup contains the same amount of its original contents.

Every year, most of the top U.S. math majors in college take the Putnam Exam, which is a twelve-question, six-hour exam. This exam is intended to weed out the very best, most promising young mathematicians, and history shows that it is sufficient, if not necessary, to score in the top ten on the Putnam to have a productive career in math. For example, Feynman scored in the top three (in fact, he scored number one, although this was not published).

Over the years, the Putnam Exam has evolved in the direction that I am trying to take the Ultra Test. Although the Putnam requires too much specialist knowledge to be an intelligence test, I reproduce below some questions from the Putnam that do not require much specialist knowledge and that I think give the feel of the exam.

- **1966 A-6:** What is $\sqrt{1+2\sqrt{1+3\sqrt{1+4\sqrt{1+\ldots}}}}$?
- **1967 A-3:** If $f(x)=nx^2-bx+c$ has integer coefficients, what is the least value of a such that $f(x)$ has two distinct zeros in $0<x<1$?
- **1972 B-2:** A particle moving on a straight line starts from rest and attains a velocity v_0 after traversing a distance s_0 . If the motion is such that the acceleration was never increasing, find the maximum time for the traversal.
- **1947 11:** aa, bb, cc, dd are distinct integers such that $(x-a)(x-b)(x-c)(x-d)=0$. If x is an integer, what is it?
- **1949 B-3:** If any two points on a closed plane curve are no more than one unit apart, what is the radius of the smallest circle that completely contains the curve?

Earlier in this issue, Kevin Langdon argues that we should submit our proposed (and supposed) intelligence tests to the peer-reviewed publication process, and proposes that we start a journal that will include academic psychometricians on the mailing list. This is an interesting non sequitur, although I think we all understand why Kevin makes it. The truth is that no reputable psychometrics journal would publish an analysis of any of our tests. This is for a variety of reasons with which we are all familiar, and which I will not belabor here.

The point I want to make is that the road to scientifically accurate and generally accepted high-level intelligence tests will be a long and winding one. I suspect that the initial inroads, by the way, will not come from psychometricians, or even from psychologists of any stripe, but rather from computer scientists, who will be working from much the same motivation that I am.

However, a journey of a thousand miles starts with a single step. We are already a few steps into this journey, and I would like to make the Ultra Test the next step. So please be on the lookout for good problems and send them in.

Rick Rosner, Editor's Comments

2024-08-22

Original publications here, December, 1994.

Editor's comments: The concept of IQ itself is slightly obsolete and ridiculous. IQ testing has a history of unsavory agendas. The arena of superhigh IQ-ology is even more problematic. The problems lie in these areas:

- **Lack of real-world performance by superhigh IQ people** (Nobel Prizes, etc.), often coupled with social awkwardness, which further reduces credibility among people who do have social skills.
- **Lack of a real-world reason to measure IQs above 150.** As most of you know, the concept of IQ was introduced to make sure education met the needs of children with varying abilities by determining whether a student has low, medium, or high ability. Schools are equally ill-equipped to meet the needs of a kid with a 150 IQ and a kid with an IQ of 170.
- **Lack of acceptance in the psychometric community** and lack of unassailable norms for superhigh IQ tests.
- **The ordeal of taking a superhigh-ceiling test,** which eliminates qualified candidates who are busy doing something other than taking IQ tests.

- Rule-bending and the dissemination of high-ceiling test answers.
- **The possibility that, in the higher reaches of IQ, IQ is inherently indeterminate**—that no number can be assigned, that no well-ordered relationship exists among high-IQ people.

I'd like to think that high-IQ people, keeping all that in mind, could treat the whole high-IQ thing with, I dunno, some lightness. Indeed, the Mega Society recently celebrated its ten-year anniversary, and only in the past few months has the issue of qualification been the site of real teeth-grinding contention. (And the recent contention resembles professional wrestling as seen through the eyes of fans who think that pro wrestling is real.)

The high-IQ world is fraught with ludicrousness, but so is everything-religion, science (The Copenhagen interpretation is pretty goofy.), any -ology. I'm sure people have been admitted to Mega through a combination of characteristics, especially persistence, augmenting less than one-in-a-million intelligence, but I doubt anyone will gain admission through unrelenting attacks and the complete destruction of an admittedly imperfect but reasonably efficient (and probably the only practical) admission system.

Let me respond to specific points:

- As far as I know, Ron Hoeflin has consistently asked to be included in the Mega Society as the founder, not as a member. He has never claimed to qualify, though I think that the general feeling among members is that he is on a par with the members.
- The Mega Society is actually a combination of two merged societies, one of which was, for part of its existence, a 1-in-100,000 society, largely because of fluctuating norms to the Mega Test. One-in-a-million is certainly a catchier cutoff, but beyond that, I don't think anyone would be too concerned about a change to one-in-a-hundred thousand. On the other hand, with all the varying norms flying about, I'm completely unpersuaded to alter the agreed-upon theoretical cutoff of one-in-a-million.
- No member will be booted out of the Mega Society or required to requalify. We suggested requalification a long time ago, and people were rightfully furious.

To get even more specific:

- While Paul Maxim's analyses of Langdon and Hoeflin and their tests have a patina of objective analysis, most readers get the impression that they seethe with resentment predating his first submission to *Noesis*. I like any material that generates responses from other people, which Maxim's material certainly does, but I don't like the distress it causes me and seems to cause others.
- The history of *Noesis* is, to a large degree, the history of Chris Langan's presentation of CTMU as a guide to the solution of Newcomb's paradox and myriad other problems, and the sometimes-surly communication between Langan and other readers. Robert Hannon's material has also generated a lot of frustrated letters.

- But both the Langan and the Hannon interactions seem to be conducted with more charity than the Maxim interactions, which make me fear for the continued existence of Mega.

Bob Williams, High Range IQ Tests – Are They Psychometrically Sound?

2024-08-22

Original publications here, February, 2021.

Preface: This manuscript [written in September 2020 - Ed. Note] is intended as a reply to questions relating to two articles in the journal *Noesis* (Mega Society journal). The questions to me were to simply ask for my thoughts on a couple of journal articles relating to tests that purport to measure IQ at ranges above those considered by professional IQ tests. I decided to render the "long" answer that is found below. I tried to cover the things that I have considered (over a period of years) relative to such tests. The answer has not been scrubbed for appearance or even the order of comments. These are simply my thoughts, with some mention of the various sources that have influenced my thinking about this category of tests.

Tests that claim IQ measurements at very high levels go by a number of names. I have traditionally called them hobby tests, because they are not designed by professionals and marketed by the companies that are dedicated to test instruments. These tests are also sometimes called high end tests, high range tests, power tests, and experimental tests. I have selected the last of these, as I think there is little inference associated with that terminology.

After some thought, I have decided to put my conclusions before the body text. I don't want anyone to be led to believe that I hold positions that might be implied, but incorrect.

Conclusions

- There are obviously people who have cognitive abilities that are above the ceilings of professional tests.

- Difficult tests can identify individuals who have very high intelligence. I doubt that anyone would argue that the Putnam Competition does not identify such people. In fact, it may be one of the very best detection vehicles available, despite its necessity for mathematics understanding. The obvious problems are that it works forsome, but not for others and it is not scaled to report IQ.
- People with very high ability may be missed by a typical experimental test, for reasons of test item weighting that impacts broad abilities.
- The rarity of people at IQ levels, above those of professional test ceilings, is the very thing that is the primary obstacle to creating a credible measurement in that range.
- Scores above the ceilings of professional IQ tests are not convincing (not even close). Test designers do not know how to properly connect the scores to reliable reference points and they do not know what we should be measuring in high ranges (see my discussion of SLODR). [SLODR = Spearman's Law of DiminishingReturns - Ed. Note]

High Range IQ Tests - Are they psychometrically sound?

Professional IQ tests (PT) are typically normed over ranges of ± 2.5 SD. A few go as high as + 3 and fewer still to + 4. Various tests have extended scoring ranges that are basically extrapolations. Hobby test designers have produced "high range" or "experimental" tests that claim very high ceilings. Many such tests are available on the internet and seem to be as popular as video games among some youngsters. These tests raise a number of topics and issues:

- Understanding fuzzy science.
- Is the norming method used valid?
- Do these tests show both internal and *external* validity?
- Is the self-report "feature" of norming valid?
- Are the tests consistent with the expected design features of professional tests?
- What should be measured? Psychometric *g*, or group factors? SLODR is the issue.
- Is the test invariant with respect to the most salient groups?
- Very long time limits versus long time limits.
- Have factor loadings been balanced?
- Measuring basic cognitive abilities, versus complex problem solving.
- Have the statistical considerations been examined by a real expert?
- Alternatives to meet the needs of exclusive clubs.

Understanding fuzzy science.

The whole discipline of cognitive science falls well outside of hard science (physics and chemistry and relatedstudies) and becomes a mixture of hard science tools (brain imaging and DNA analysis) and fuzzy science applications. We already know this, but for review, we are

measuring intelligence, using an equal interval scale that isproduced, by manipulation, from an ordinal scale and then centered on an average of a small group of people that can vary from test to test, by calendar date, and between nations.

The impressive thing is that, after all of these imprecise maneuvers, we actually can get test results that can be shown to be meaningful and predictive of life outcomes. We must not lose sight of the fact that our measurementtechniques are a set of really fuzzy methods and are not in the category of measuring something with an interferometer or micrometer. Even in hard sciences, the instruments used tend to be accurate over a limited range and either stop or become distorted beyond those ranges, so people don't make claims that a measurement is accurate when they know their peers are quite aware of what the instruments can deliver.

In the case of experimental tests we are operating in a range where the instruments in question have not been well-calibrated and are not used by professionals and are even probing into areas where the things being measured may be different from what is found over more than 99% of the range of the parameters. [One-in-100 IQ ~135 and one-in-1,000 IQ ~147 -Ed. Note]

Is the norming method used valid?

Various experimental tests designers presumably use different methods of linking scores to what they believe are meaningful IQs. I cannot claim to have seen or evaluated these approaches, but I doubt that they generally conform to the methods used by PT designers. Assuming Classical Test Theory (more about that later), the designers effectively force fit a Gaussian distribution to the results of tests given to the norming group. This means that the test is dependent on the norming group being large enough and representative of the group the test is intended to serve. The force fitting is done by adding and subtracting test items with difficulty levels that will increase or decrease the number of correct answers at the points on the distribution which do not fit the normal curve.

If the test author simply selects a group of test items and uses the whole lot, he is not likely to end up with a good fit.He needs to start with extra test items and select only those that produce a good fit to the Gaussian distribution.

Here we again meet fuzzy issues. For starters, we have no idea how IQ is distributed beyond the ceilings of well designed, comprehensive PTs. The experimental test designer knows that he is not dealing with a full range of data,so he is trying to fit to the right tail, not really knowing where he is on the tail, nor its real world shape.

At this point, we should go to chapter 4 of A.R. Jensen's *Bias in Mental Testing* (1980). New York: Free Press. It is a discussion of why it is reasonable to assume a Gaussian distribution. Jensen makes the point that the distribution is indeed normal over the ranges of PTs and does so in thousands of words. One illustration goes back to my comments about the fuzzy nature of what we are doing:

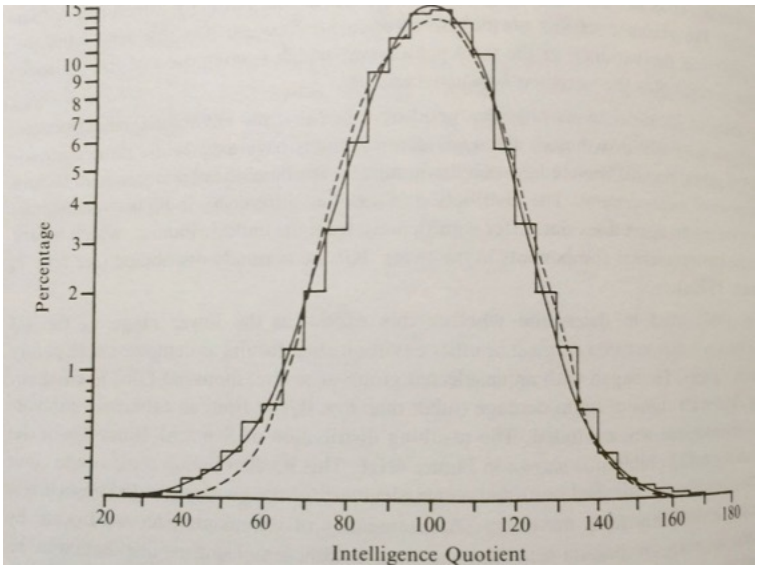


Figure 4N.1. (above) Distribution of Stanford-Binet [S-B]IQs of a sample of 4,523 London children from which all cases of diagnosed brain damage and extreme environmental deprivation have been excluded. A normal curve (dashed line) and Pearson's Type IV curve (continuous line) are superimposed on the actual data (stepwise curve). Note that the Type IV curve shows a closer fit to the data than does the normal curve. (From Burt, 1963, p. 180) [Jensen, *Bias in Mental Testing*, page 120 -Ed. Note]

First, note that the S-B results in the figure (reasonably large N) are a closer fit to a Type IV curve than a normal curve. We use the Gaussian distribution because it works well enough, given the fuzzy nature of otherconsiderations;; because it simply doesn't matter (look at the two fit options);; and because Mother Nature generallydoes things that fit a normal distribution. Here is what Jensen [on page 88 of *The g Factor* -Ed. Note] wrote about the fit:

There are plausible reasons, however, for assuming that individual differences in *g* have an approximately normal, orGaussian ("bell-shaped"), distribution, at least within the range of ± 2 σ from the mean. That range is equivalent to IQs from 70 to 130 on the typical IQ scale (i.e., μ = 100, σ = 15). Individual differences in general mental ability are usuallymeasured by some test or combination of tests that are highly *g* loaded, and such tests are purposely constructed to have an approximately normal or bell-shaped distribution in the standardization population. Although the normal distribution of test scores is usually accomplished by means of certain psychometric artifices, it has a quite defensible rationale.

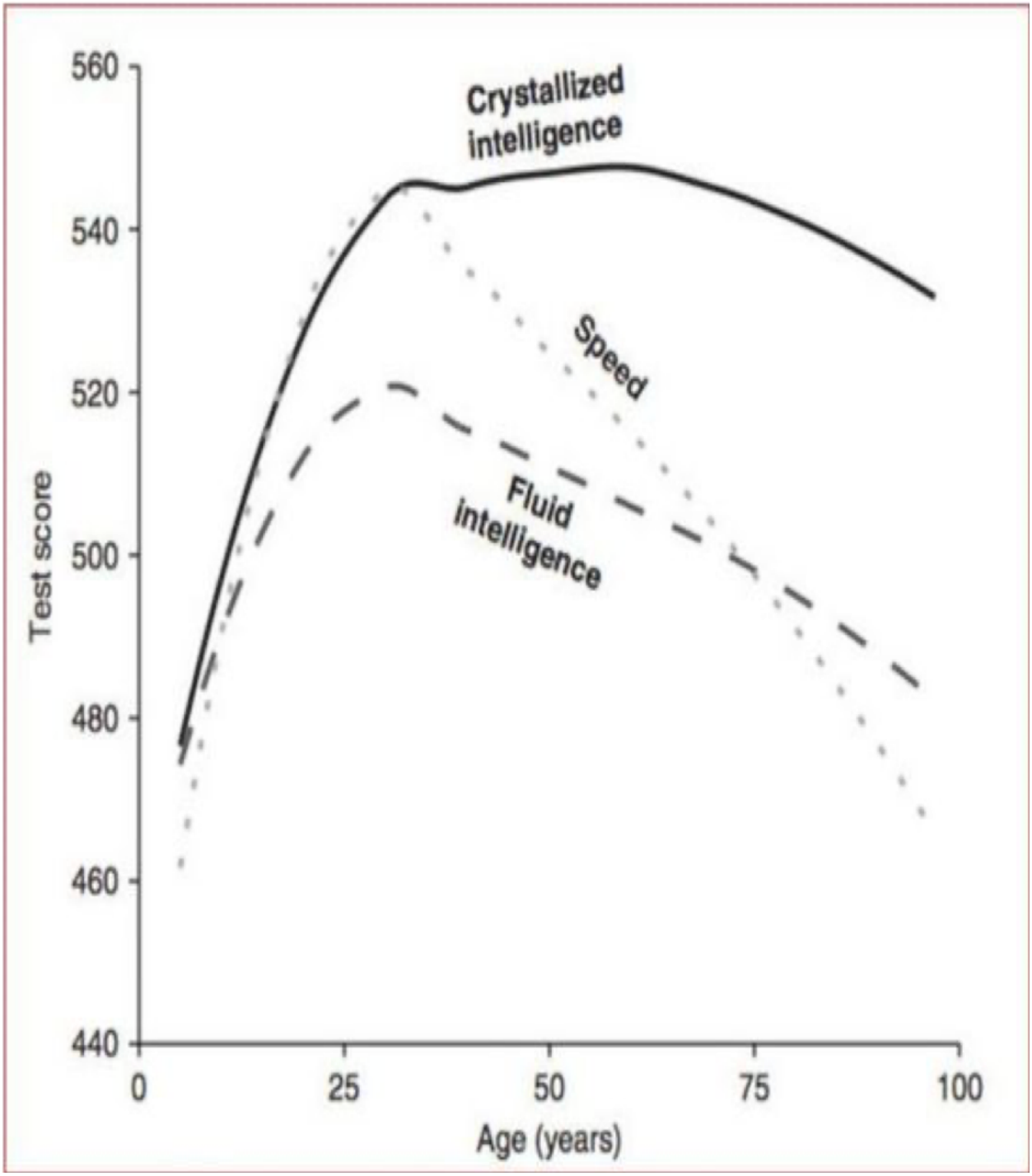
Covering a range of ± 2 is a much more reasonable claim for the fit than trying to extend it by several additionalstandard deviations. We use the curve over a limited range because it works, not because it is the direct cause of intelligence over any range. My point here is to again point out that we are in a fuzzy world, not dealing with hard science.

Do these tests show both internal and *external* validity?

When a PT is produced, it is evaluated against a wide range of statistical measures that indicate whether the test does what it is supposed to do. First, is internal (construct) validity, which

simply means: does the test measure the thing it claims to measure. The methods used to establish construct validity are messy, compared to some others and are discussed in *Bias in Mental Testing* [e.g., pages 303-305 -Ed. Note]. The test can be factor analyzed and compared to the factor analysis of an accepted standard test (Wechsler, etc.) and can be otherwise directly compared to tests that have historically produced consistent results.

Caption for image (below): Accomplishments across individual differences within the top 1% of mathematical reasoning ability 25+ years after identification at age 13. Participants from Study of Mathematically Precocious Youth (SMPY) Cohorts 1, 2, and 3 (N = 2, 385) are separated into quartiles based on their age - 13 SAT-M score. The quartiles are plotted along the x-axis by their meanSAT-M score. The cutoff for a score in the top 1% of cognitive ability was 390, and the maximum possible score was 800. Oddsratios (OR) comparing the odds of each outcome in the top (Q4) and bottom (Q1) SAT-M quartiles are displayed at the end of every respective criterion line. An asterisk indicates that the odds of the outcome in Q4 was significantly greater than Q1. STEM = science, technology, engineering, or mathematics. STEM Tenure (Top 50) = tenure in a STEM field at a U.S. university ranked in the top 50 by U.S. News and World Report's "America's Best Colleges 2007." Adapted in part from Park, Lubinski, and Benbow (2007, 2008).



The real issue is external (predictive) validity. This is the heart and soul of any IQ test. If the test does not predict real world outcomes that are independent of the test, there is no practical use for the test. Readers here already know the factors that are most strongly correlated with IQ, so there is no need to explain what needs to be checked. In the case of this discussion, we are concerned with experimental tests and whether they are or are not valid. I submit that the acid test is external validity. If we go to the high end of the spectrum (our reason for this discussion) we have an excellent demonstration of the predictive validity of intelligence for the SMPY cohorts.

The question that must be convincingly asked of experimental tests designers is whether they can show similarexternal data, verifying that there are differences in such things as educational achievement, income, publications, etc. that are predicted (presumably increasing) by the IQ scores these tests report above the ceilings of PTs. If the answer is NO, what is the reason for the tests? This is the meat and potatoes of IQ testing. Predictions must be verifiable.

Is the self-report "feature" of norming valid?

Among the fuzzy aspects of attempting to design a measurement tool for high IQ is the use of self-reports. I seeseveral concerns with such data playing a vital role in norming. If the self-reports are from different tests (as one wouldassume to be the case), even flawless self-reports are from tests that share a relatively small amount of variance. The data below are old, but indicate typical test-A to test-B correlations:

- WAIS to Stanford Binet = 0.77
- WAIS to Raven's = 0.72
- WAIS to Otis = 0.78
- WAIS to SAT = 0.80

Seligman, D. (1994) A Question of Intelligence: The IQ Debate in America. New York City. Carol Publishing Group. (Page 167)

The first comparison is a 59% shared variance. But we all know that self-reports have a degree of error. Although it is not a direct comparison, when people are asked to estimate their IQ, the result is a staggering overstatement of 30 IQpoints! (I understand that self-estimate and self-reporting are different. But I think this observation is worth keeping in mind.)

(Gilles E. Gignac, Marcin Zajenkowski, People tend to overestimate their romantic partner's intelligence even more than their own, *Intelligence*, Volume 73, 2019, Pages 41-51.)

The issue of self-reports has been studied extensively, with results somewhat dependent on the nature of the information being reported. Here is a reasonable comment on the general topic:

Cook and Campbell (1979) have pointed out that subjects (a) tend to report what they believe the researcher expects to see, or (b) report what reflects positively on their own abilities, knowledge, beliefs, or opinions. Another concern about such data centers on whether subjects are able to accurately recall past behaviors. Cognitive psychologists have warned that thehuman memory is fallible (Schacter, 1999) and thus the reliability of self-reported data is tenuous.

Source: Chong-ho Yu (2013) Reliability of self-report data.

It is understandable that a designer would want to use self-reported test scores in an attempt to link his test to a PT, but the number of error sources is obviously large – to the point of making such an effort seem futile.

Are the tests consistent with the expected design features of professional tests?

For any comprehensive IQ test to function properly it must have test items of different levels of difficulty. I am unaware of the methods used by experimental test designers, but if they do not have a reliable means of determining relative difficulty, they are shooting in the dark.

From chapter 4 of Bias in Mental Testing:

“The simple fact is that a test unavoidably yields a near normal distribution when it is made up of (1) a large number of items, (2) a wide range of item difficulties, (3) no marked gaps in item difficulties, (4) a variety ofcontent or forms, and (5) items that have a significant correlation with the sum of all other item scores, so as to ensure that each item in the test measures whatever the test as a whole measures. (Items that are uncorrelated or negativelycorrelated with the total score can only add error to the total scores.) These are all desirable features of a test.”

In the case of experimental tests, my impression is that these things pose problems. Is the number of items highenough when the test is intended to be very difficult and worked on for weeks?

Does the designer know the item difficulty of each item? If he does not know, how does he conclude that he has not created large gaps in difficulty or bunched many test items together because they have the same item level difficulty?

Are the items sufficiently diverse with respect to Jensen’s 4th condition (content variety)? In constructing a comprehensive IQ test, professional designers repeatedly note that the test must be diverse. Here is what differentiates PTs of different quality:

Poor = 1 Fair =2 Good = 3 Excellent = 4

1. Number of tests	1	1-2	2-8	9
2. Dimensions	1	1-2	2-3	3
3. Testing time	3-9 min	10-19 min	20-39 min	40+ min
4. Correlations to *g*	⩽ .49	.50-.71	.72-.94	⩾ .95

Source: Gilles E. Gignac, Timothy C. Bates (2017) “Brain volume and intelligence: The moderating role of intelligence measurement quality”;; *Intelligence* 64, 18–29.

Clearly, the number of tests (subtests) needs to be 9 or more in order to produce at least three second order factors,from which *g* can be extracted. Do experimental tests really use that many subtests? In my opinion, the number of dimensions should be higher for a comprehensive test.

The real problem here is that we are measuring in the range in which we have no way to know what is happening to *g*,other than it is likely that SLODR is a serious consideration and that the *g* variance is no longer a linear indicator of intelligence.

When an experimental test is used, the designer may have to deal with some difficulty in calculating the reliability coefficient. It is simply the ratio of the variance in the true test score, divided by the error variance. The methods used to calculate the reliability coefficient are typically to effectively administer the test twice (by designing an almost identical test, with test

items of equal levels of difficulty) or by using the split-half method (correlating the two halves, then applying the result to the Spearman-Brown formula. If the number of test scores available is small, the process is difficult. I have not seen any indication that designers have used double testing of the same group, which leaves only the split-half method. Hopefully there is data showing that this methodwas properly applied to each experimental test and that the results show a coefficient close to 0.90.

Most IQ tests have historically been designed to be scored using Classical Test Theory. But, since the development of Item Response Theory [IRT -Ed. Note], we have a clearly superior method that is based on known item level difficulty.IRT still requires that the test items be given to a reference group, which matches the characteristics of the full population of the group to be represented by the finished test. The developer determines an Item Characteristic Curve for each test item and can then determine IQ based on the item difficulty of the most difficult items that have correct responses from the testee. This method is particularly well suited to use in computer testing, since the computer canpresent more or less difficult test items in response to correct or incorrect responses. In the case of experimental tests the number of test items may be too small for IRT to be practically applied, but if it were used, at least the designer could avoid using test items that are too close to the same level of difficulty and space the items, such that there are not large gaps in the levels of difficulty. If item level difficulty is not known, it is difficult to believe that anything other than luck would make the test perform properly – particularly over a very high intelligence range.

What should be measured? Psychometric *g*, or group factors? SLODR is the issue.

IQ tests measure *g*, non-*g* residuals of broad abilities, and uniqueness. The sum of the variances of these must equal 100%. Over most of the IQ spectrum (let's think in terms of ± 2 SD for discussion), the thing we are looking for(outside of clinical applications) is *g*, because it accounts for virtually all of the predictive validity of the test. We usethe IQ test as a proxy for *g* because there is usually enough *g* saturation to justify the proxy.

When high levels of intelligence are involved, there is the possibility that *g* is not following a linear increase with the measurement from the test in question. This is Spearman's Law of Diminishing Returns (SLODR). Jensen wrote (seeAppendix A - *The g Factor*): "The higher a person's level of *g*, the less important it becomes in the variety of abilities the person possesses."

Evans explained the situation well: "The possibility of a breakdown of *g* at higher levels of intelligence, even with a narrow range of tests (as in the Armed Services Vocational Aptitude Battery) implies that we may have to reexamine the nature of intelligence." … "There may be a single driving factor at low levels of *g*, but this may be manifested in a variety of different ways at high levels of *g*." (Evans, M. G. (1999). "On the asymmetry of *g*," Joseph L. Rotman Schoolof Management, University of Toronto.)

A clean proof of exactly what is happening is difficult. There has been about one paper addressing this at most ISIR conferences over the past decade or more. In 2004, I asked Jensen if it would be possible to calculate *g* in relatively narrow slices to prove the effect. He confidently replied that comparing the top and bottom halves (which had already been done) was about the

best that could be done. The various papers I mentioned have approached the problem from different directions with mixed results. My impression is that the weight of evidence is that Spearman was right and SLODR is real. Unfortunately, that does not tell us if it is a small, moderate, or large effect, nor does ittell us how much the effect size might change at very high levels.

In a qualitative sense, we see that bright people seem to obviously demonstrate that there are increasing differencesin their areas of highest performance. Jensen discusses this qualitative observation in various places, including the Appendix A, mentioned above.

When a test seeks to measure at very high levels, it is dealing with intelligence that is not structurally the same as that found at lower levels and which seems to become particularly characterized by non-*g* factors. Do experimental teststake this into account? If they claim to do equally well as a PT below the usual ceilings, what suggests that they also do well when the nature of intelligence is different?

Let's consider a test-A that consists of mostly verbal test items and another, test-B, that consists of many spatial items.If we are attempting to measure IQ at very high levels, it is very likely that a person who scores well on test-A will notscore well on test-B (or vice versa). How do we rationally compare someone with high and narrow verbal ability to someone with high and narrow spatial ability? Isn't this similar to comparing the abilities of a superb painter to those ofa world class mathematician? My perspective is that when SLODR kicks in the utility of IQ tests is damaged because we rely on a linear *g* throughout the major range of interest, then we try to use the same measurement approaches when intelligence biologically changes. It is much like studying a solid and understanding its thermal expansion until the temperature causes the solid to start melting.

Is the test invariant with respect to the most salient groups?

PTs are typically evaluated to show that they are invariant relative to race, sex, and occasionally other groups (age is a special case). If the tests are designed around data that is skewed toward a racial or ethnic group, is the test valid for other groups? If invariance is not shown, there should be a cautionary flag that the test is only for specific listed groups. Invariance is typically shown by multiple group confirmatory factor analysis. If invariance is not established, for race/ethnicity (example), the interpretation of the results of testing a group may have an embedded error of unknown magnitude. Does the test only work for one sex, or has invariance by sex been confirmed?

In the case of age, there are several considerations. Does the scoring determine z-scores by age groups, or are theyall taken as the same? (Age adjustment is a necessary requirement for a professional IQ test, unless the test specification is limited to a narrow age range. IQ is as defined by David Wechsler's method and is relative to age peers.) We know that the brain is constantly changing from birth to death. One of the things that we accept as a given is that IQ measurements are relative to age peers. The figure below relates directly to this concern.

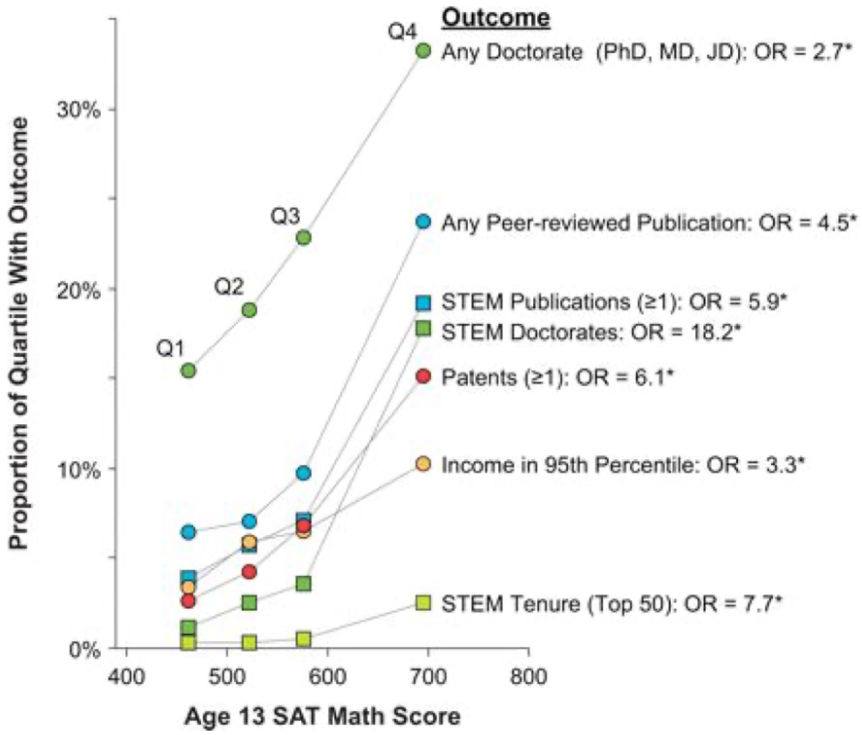


Figure source: *Intelligence: All That Matters*, S.J. Ritchie, John Murray Learning, London (2015).

It is vital that age effects are taken into account or the test will not reflect the thing we understand as IQ. The differences between age 30 and age 50 are large. Naturally invariance must be established for age. If it is not, the test should specify the age range over which it is known to be invariant.

Establishing sex invariance could be even more challenging than simply finding an appropriate norming group. The combined factors of a higher male mean IQ (age 16 and above) and the higher male SD results in a huge difference of males and females at high levels of intelligence. For example, data for whites:

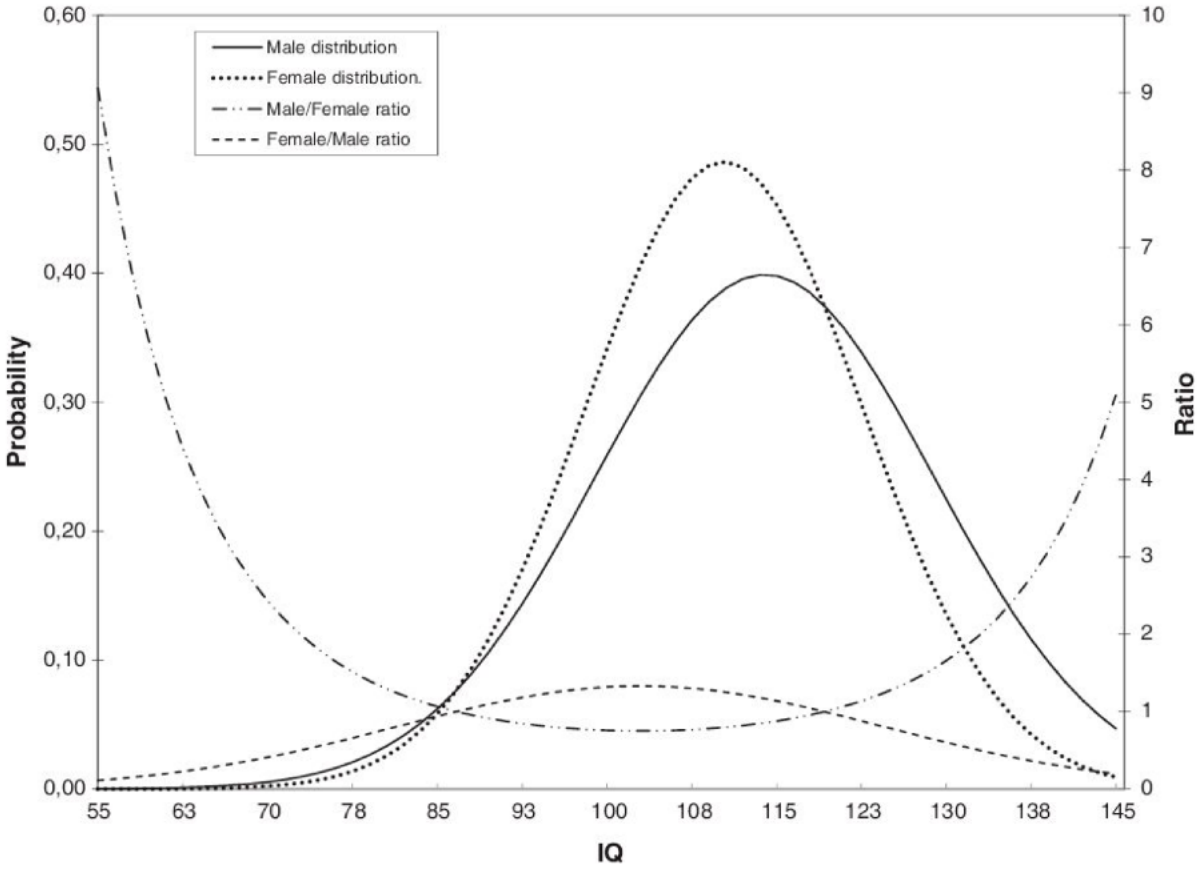


Figure source: Sex differences across different racial ability levels: Theories of origin and societal consequences;; Helmuth Nyborg;; Intelligence 52 (2015) 44–62.

The sex ratio at IQ 145 is already 7 males per 1 female and is increasing rapidly. Interestingly, the ratio is higher for Hispanics and higher still for blacks.

Very long time limits versus long time limits.

Let's forget SLODR now and look at other features, as used in PTs. Many of the test items that are used in PTs can be given without time limits. It is well established that removing time limits does not reduce the reliability of these items, nor does the use of adequately long time limits. The problem that arises is that tests without time limits that are actually expected to take weeks to complete cannot measure any of the factors associated with time. Mental speed accounts for up to 80% of the variance in intelligence (there is a lot of covariance, particularly with WMC). [WorkingMemory Capacity = "The capacity to selectively maintain and manipulate goal-relevant information without getting distracted by irrelevant information over short intervals."

Source: https://www.sciencedirect.com/topics/psychology/working-memory-capacity -Ed. Note]

Giving up this whole category of measures significantly reduces the subtests that can be used.

It gets worse…Working Memory Capacity (WMC) also accounts for up to 80% of the score variance (obviously this means a lot of covariance with mental speed). The literature is stacked

with studies showing very high relationships between WMC and various parameters related to intelligence (*g*, Gf, executive function, attention, inhibition, etc.). While it may be argued that these are indirectly measured by very difficult test items, skipping the step of direct measurement seems to be a detriment to the strength of a high quality test. The simple fact is that WM measurements are time dependent.

(There is an indirect speed effect, in other measures, due to the Gs correlation with Gf and WMC.)

[https://en.wikipedia.org/wiki/Cattell%E2%80%93Horn%E2%80%93Carroll_theory -Ed. Note]

Have factor loadings been balanced?

The *g* loading of a subtest is somewhat dependent on the structure of the other parts of the test. Over or under representation of factors being tested directly influences the other test items. For this reason PTs seek a balancebetween the numbers of test items in a given test category. Can we depend on the experimental tests to be balancedin its structure? If not, one might argue that Spearman's indifference of the indicator works in favor of the narrow test (RPM is an example), but this is the very point that differentiates a high quality test from a lesser one (see the previous figure from Gignac and Bates). The problem is that experimental tests show deficiencies in many areas that are important to PTs.

[RPM = Raven's Progressive Matrices, "a nonverbal test typically used to measure general human intelligence and abstract reasoning" thereby gauging fluid intelligence.

Cf. https://en.wikipedia.org/wiki/Raven%27s_Progressive_Matrices -Ed. Note]

Measuring basic cognitive abilities, versus complex problem solving

When I looked in Intelligence to see how many papers were about complex problem solving, I found that the numberis very high (so high I don't believe it). Of the papers I have read, I have generally been impressed that they show littledifference between testing people with complex problems and testing them with a comprehensive test, such as the WAIS or Woodcock-Johnson.

I previously mentioned SLODR, but I would point out that there are also diminishing returns in testing. As JamesThompson has pointed out in at least 3 articles on The Unz Review, there are some very quick (as in 2 minutes or less) tests that do quite well when correlated with high quality PTs. This is a wonderful example of diminishing returns. As the test used is improved with more test items, more factors, longer testing time, etc. there are gains in reliability, and presumably various things such as internal and external validity, ceiling, floor, and *g* loading, but these come in ever-decreasing increments. I am trying to illustrate that an experimental test is attempting to make a large move in the direction of ceiling and that achieving such a change should come at a very high cost.

It is my general impression that experimental tests are intended to probe some traditional second order factors by using test items that have high difficulty. To the extent that these are multi-step solutions, they probably can be fairly classified as complex problem solving test items. In that case, there is a huge amount of literature to evaluate, if the designer is concerned about changing the nature of IQ testing to CPS [complex problem solving -Ed. note] format. It may be

justifiable, but should be understood by the designer, in terms of how it actually impacts test results and theirrelation to standard PTs. For example, Alexander Christ, et al. (*Intelligence* 78 (2020) 101421) noted that complexproblem solving and reasoning have proved to be distinct constructs.

Have the statistical considerations been examined by a real expert?

The question above is self explanatory. For example, has the author of the experimental tests designed conventionaltests that are in use by clinical psychologists, intelligence researchers, and which are referenced in the major textbooks on this subject? If not, it would seem to be helpful to have a person with such a background evaluate the test in the various categories I have mentioned.

I am aware that some experimental test designers are quite intelligent, well educated, and believe themselves to befully competent to design an outstanding IQ test. Those people should, in my opinion, meet the simple standards I listed.

Alternatives to meet the needs of exclusive clubs

The problem that I see with experimental tests is that they purport to actually measure IQ in a range that is beyond the reasonable ceilings of PTs. One option is to not claim that the tests measure IQ and simply give a raw score. Allow clubmembership based on the raw score alone, without attempting to connect the score to IQ.

It may be possible to use tests given in early life (as is done with SMPY), if the rarity of the score can be established. I have doubts that much can be done beyond the published 1 in 10,000 results from the Vanderbilt researchers.

As I see it, a primary problem is that the design of experimental tests is unlikely to ever be done with a full range reference group (from very low to very high). The designers try to invent a way to work on only a piece of the right tail.The problem is that they don't know how to set the data, such that it actually fits the extreme high end of a Gaussian distribution. I honestly think this problem is not going to be resolved in a convincing way, using existing methods. In his discussion of how to best deal with researching the Flynn Effect, Jensen suggested anchoring test scores to biological measures:

"As I have suggested elsewhere, conventional psychometric raw scores will need to be anchored to measures that presumably are not influenced by the environmental variables that raise test scores without increasing *g*. The anchor variables would consist of measures of reaction time to various elementary cognitive tasks, evoked brain potentials, nerve conduction velocity, and the like, that are demonstrably *g*-loaded. (A composite measure based on the anchor variables should have a reasonably high correlation [say, $r > .50$] with the psychometric test scores.) Mental test raw scores would be regressed on these anchor variables in a representative sample of some population.

…That is, the mean gain would be reflected in the anchor variables as well as in the test scores."

It might be possible to find biological measures that could be used to anchor a common point between a test such asthe WAIS and an experimental test and then have some confidence that the

joint is based on something measurable. Unfortunately, simply finding a common point does not resolve the other issues that strike me as murky, such as theissues associated with SLODR and general test design considerations. The comments directly below may offer more promise.

We live in a rapidly changing world. Only a few years ago, brain imaging was very limited, but when MRI technology was introduced, there was a sudden explosion of living brain data that had not previously been seen. Neurologists are turning out studies faster than anyone can reasonably read them. As far back as 2006, Richard Haier told the ISIR conference that we would be able to measure intelligence entirely from brain imaging and that it was not too far off. I later asked him for more insight on it and he told me that the problem was basically that – in 2007 – it takes one of the few expert researchers to read the data. This makes the cost too high to be practical. Of course, the "reading" will be automated and probably be enhanced by using machine learning. The most likely method will be one that combines many measurable factors (cortical thickness, cortical surface area, size and shape of the corpus callosum, fractional anisotropy of gray and white matter, and measures of the connective networks (such as mean path length and numbers of connections to major and minor hubs). We already have a patent that is limited to CT: **Patent US8301223** - Neurobiological method for measuring human intelligence and system for the same.

"The method enables neurometric IQ to be measured by processing MRI and fMRI images of a subject to determine cortical thicknesses and brain activation level, determining structural IQ (sIQ) and functional IQ (fIQ) from the determined cortical thicknesses and brain activation level, and using the structural IQ (sIQ) and the functional IQ (fIQ) as predictors to measure the neurometric IQ of the subject. With this, individual differences in general cognitive ability can be easilyassessed. It suggests that general cognitive ability can be explained by two different neural bases or traits: facilitation of neural circuits and accumulation of crystallized knowledge."

Imaging technology may eventually create very high ceilings – or not. The obvious problem is that we must wait forthis to be fully developed and automated to the point of reasonable costs. Haier told me that he envisioned a full psychometric evaluation in 20 minutes at a 2007 cost of about $200.

Christopher Harding

2024-08-22

Original publication 1 and 2.

Christopher Harding is the Founder of the International Society for Philosophical Enquiry (ISPE), and a Member of OlympIQ Society and the ESOTERIQ Society. He was born on August 4, 1944 in Clovelly Private Nursing Home at Keynsham, Somerset, English, United Kingdom. He has never married. He arrived in Rockhampton, Queensland, Australia, in the morning of October 11, 1952. He remains there to this day. He has held memberships with the Eugenics Society (1963-1964), the British Astronomical Association (1964-1969), the International Heuristic Association (1970-1974), the Triple Nine Society (1979-1990 & 1992-1995), the 606 Society (1981-1982), the Omega Society (1983-1991), the Prometheus Society (1984-1990), the International Biographical Association (1985-1990), Geniuses of Distinction Society (1986-1988), the American Biographical Institute Research Association (1986-1990), the Cincinnatus Society (1987-1990), the 4 Sigma Group of Societies [incorporating all groups having 4 Sigma plus cut off points] (1988-1990), The Minerva Society [Formerly the Phoenix Society] (1988-1990), The Confederation of Chivalry (1988-1990), the Planetary Society (1989-1990), Maison Internationale des Intellectuels [M.I.D.I.] (1989-1990), TOPS HIQ Society (1989-1990), the Cleo Society (1990-1991), the Camelopard Society (1991-1992), the Hoeflin One-in-a-Thousand Society (1992-1993), the Pi Society (also like the Mega Society for persons

with 1 in one million I.Q. level (5th April 2001 – 2002), INTERTEL [The International Legion of Intelligence] (June 1971-March 2010), The Hundred (1972-1977), the New Zealand National Mensa (1980-1982), and the Single Gourmet (1989-1991), among numerous other memberships, awards, and achievements (For the most recent or up-to-date information, please see the ESOTERIQ Society listing: https://esoteriqsociety.com/esotericists/esoteriq-id06/); growing up; a sense of the family legacy; the family background; experience with peers and schoolmates; qualifications; purpose of intelligence tests; high intelligence; the geniuses of the past; the greatest geniuses in history; a genius from a profoundly intelligent person; profound intelligence necessary for genius; some work experiences; job path; the idea of the gifted and geniuses; the God concept or gods idea; science; some of the tests taken; the range of the scores; ethical philosophy; social philosophy; economic philosophy; political philosophy; metaphysics; philosophical system; meaning in life; meaning; an afterlife; the mystery and transience of life; love; *National Enquirer;* the gap between cognitive abilities and record of employment; living situation without a record of work; alone; the professionals test someone just shy of 1-year-old; parents react to being called "liars to their faces"; genius; intelligence tests; publications or periodicals; artificial constructs; the factors making genius; God as human idealism; the Concept of God; science; the areas most affected by this despoilment; the areas least affected by this despoilment; 6-sigma; the ESOTERIQ Society; conclusions; and the information in Quantum Physics.

Scott Douglas Jacobsen: When you were growing up, what were some of the prominent family stories being told over time?

Chris Harding: Where we came from and who we were.

Jacobsen: Have these stories helped provide a sense of an extended self or a sense of the family legacy?

Harding: They were depressing as I could not live up to them.

Jacobsen: What was the family background, e.g., geography, culture, language, and religion or lack thereof?

Harding: Varied. Mostly titled aristocracy and connections to Royal Houses.

Jacobsen: How was the experience with peers and schoolmates as a child and an adolescent?

Harding: Non-existent.

Jacobsen: What have been some professional certifications, qualifications, and trainings earned by you?

Harding: None.

Jacobsen: What is the purpose of intelligence tests for you?

Harding: They are something on the side.

Jacobsen: When was high intelligence discovered for you?

Harding: 2 days before my first birthday. My parents had me tested. When speaking of me they were called liars to their faces.

Jacobsen: When you think of the ways in which the geniuses of the past have either been mocked, vilified, and condemned if not killed, or praised, flattered, platformed, and revered, what seems like the reason for the extreme reactions to and treatment of geniuses? Many alive today seem camera shy – many, not all.

Harding: I recall a quote from the Journal of the British Eugenics Society. "They want the Genius, but not its loathsome owner."

Jacobsen: Who seem like the greatest geniuses in history to you?

Harding: "Leonardo, complex solitary, a Master Genius in an age of Genius." In his life, it was said of him,"It is beyond the power of nature to create another man like Leonardo," yet his final recorded words were "I have failed mankind and I have failed God."

Jacobsen: What differentiates a genius from a profoundly intelligent person?

Harding: Genius is creative ability of the highest possible kind.

Jacobsen: Is profound intelligence necessary for genius?

Harding: No, Genius implies the narrowing of intelligence.

Jacobsen: What have been some work experiences and jobs held by you?

Harding: Absolutely nothing at all.

Jacobsen: Why pursue this particular job path?

Harding: I never did.

Jacobsen: What are some of the more important aspects of the idea of the gifted and geniuses? Those myths that pervade the cultures of the world. What are those myths? What truths dispel them?

Harding: They march to the beat of their own drum.

Jacobsen: Any thoughts on the God concept or gods idea and philosophy, theology, and religion?

Harding: God is purely human idealism; largely what you can't attain. The Concept is set beyond what can be considered.

Jacobsen: How much does science play into the worldview for you?

Harding: Science has become despoiled with its obsession with consensus and ignorance of the paradigm shift. Some one point Einstein to a newspaper article "One hundred against Einstein" to which he replied "It would only take one". Less and less.

Jacobsen: What have been some of the tests taken and scores earned (with standard deviations) for you?

Harding: Several times scoring over plus six sigma.

Jacobsen: What is the range of the scores for you? The scores earned on alternative intelligence tests tend to produce a wide smattering of data points rather than clusters, typically.

Harding: Ceiling limitations were the biggest problem; in which case I could finish them well and truly before the time limit was up. For these the test was useless.

Jacobsen: What ethical philosophy makes some sense, even the most workable sense to you?

Harding: None.

Jacobsen: What social philosophy makes some sense, even the most workable sense to you?

Harding: None.

Jacobsen: What economic philosophy makes some sense, even the most workable sense to you?

Harding: Also none.

Jacobsen: What political philosophy makes some sense, even the most workable sense to you?

Harding: Also none. If you join a political group, you wind up as an apologist for them!

Jacobsen: What metaphysics makes some sense to you, even the most workable sense to you?

Harding: None.

Jacobsen: What worldview-encompassing philosophical system makes some sense, even the most workable sense to you?

Harding: None, philosophy is word juggling!

Jacobsen: What provides meaning in life for you?

Harding: There is no meaning in Dictionaries only associations with other words: Meaning in life is the same; you make the meanings.

Jacobsen: Is meaning externally derived, internally generated, both, or something else?

Harding: Meaning is only a PATTERN.

Jacobsen: Do you believe in an afterlife? If so, why, and what form? If not, why not?

Harding: The only afterlife [an oxymoron] is the "truth" in QUANTUM PHYSICS: Just as in Classical Physics energy and matter can not be destroyed only converted one into the other; in Quantum physics information can not be gained or lost, it some how just IS.

Jacobsen: What do you make of the mystery and transience of life?

Harding: It like everything else is BOUNDED. This is a condition of being defined.

Jacobsen: What is love to you?

Harding: Love is simply TRANSFERENCE [See Freud].

Jacobsen: What Royal Houses were the main connections with family?

Harding: Most prominent – French side.

Jacobsen: In the *National Enquirer* published on June 25, 1991, there was an article about a certain man with the "world's highest IQ" who is a "jobless janitor." What did this particular media attention do for you?

Harding: Nothing.

Jacobsen: I state the caveat of "absolutely nothing at all" as the response to the work experiences question. It is *reported* that you have worked in menial jobs and had stretches of unemployment, e.g., in the *National Enquirer*. What explains the gap between the cognitive abilities and the cognitive demands of the jobs for you? Alternatively, what explains the gap between cognitive abilities and record of employment for you?

Harding: Unknown.

Jacobsen: How did you sustain yourself in terms of living situation without a record of work?

Harding: Family.

Jacobsen: Why the "non-existent" life with peers and schoolmates? Did you feel alone?

Harding: Violence and exclusion.

Jacobsen: How did the professionals test someone just shy of 1-year-old? It seems odd, even stranger than the 2-and-a-half-year-old, or thereabouts, cases entering Mensa International (or their national group).

Harding: Mental age in my case 3 years 4 months made that easy!

Jacobsen: How did your parents react to being called "liars to their faces" when 'speaking of you'?

Harding: They were taken aback by this.

Jacobsen: Does this desire of cultures wanting genius while not wanting *the genius* create a toxic dichotomy in the general culture? Something to which only lip service is paid, while wanting to kill in former times, and 'kill' in modern times, the genius.

Harding: It comes from competitiveness [jack is equal to his master]. In many cultures submissiveness is considered politeness. That is considered standard in communication. It is why first world cultures see themselves as superior.

Jacobsen: As these intelligence tests have been a part of life before even 1-year-old, may I ask, what has been the life lesson from them for you?

Harding: Look, people see I.Q.'s as not valid above their own. Everybody does this. It is very noticeable that children asked who in their class is smartest will name themselves!

Jacobsen: As you recalled the quote from the *Journal of the British Eugenics Society*, I'm sure many will be interested now. What publications or periodicals do you continue to read now? What ones did you previously read and no longer do so?

Harding: No preference; I am a total generalist.

Jacobsen: With Leonardo da Vinci as "a Master Genius in an age of Genius," do you think artificial constructs could fill the gap between genius seen before and unseen genius now, i.e., artificial constructs with the capabilities of the highest human genius?

Harding: They have provided little evidence they are going to solve this one: My Mother once said the process was 'ant like' rather than a G-function.

Jacobsen: What are the factors making genius "creative ability of the highest possible kind"? Other than the qualities inherent in 'marching to the beat of their own drum.'

Harding: Genius by definition would be individualistic. As one person said to me, I was very `singular'.

Jacobsen: If "God is purely human idealism; largely what you can't attain," what are some exceptions to this thing one "largely… can't attain" or the things attainable within this definition of God as human idealism?

Harding: What I meant was the problem lay beyond the nature of logical process. It is answerable in terms of the proof of the last theory of sets. But you still get back to the conclusion that if God exists he either is the Universe or does not exist.

You are still dealing with value judgments or in assigning names; which amounts to the same thing. My Brother agreed with me that the highest form of reasoning was EVALUATION. Since to invoke reason one must first evaluate a proposition.

Jacobsen: Is the setting of the "Concept… beyond what can be considered" a defense against formal knockdown critique of the Concept of God?

Harding: No.

Jacobsen: When did science begin this despoilment with the obsession with "consensus and ignorance"?

Harding: Always was there. In our own time many people use science to moralize, and science has become the new religion. This can't be done of course. There is no bridge either between philosophy and religion.

Jacobsen: What are the areas most affected by this despoilment?

Harding: It is seen in notions of anthropomorphism with regard to climate change. Not so! The real cause is the Sun. Note, Astronomers had long ago pinned this down to Sun Spot Cycles. A new 11+ year Cycle began last year and rising temperatures have turned back. One Russian Woman Scientist predicts the onset of a period of dropping temperatures starting around 2030, though this figure is very uncertain!

Jacobsen: What are the areas least affected by this despoilment?

Harding: Human aging and Quantum Physics–much progress continues at the moment.

Jacobsen: What were the tests when scoring above 6-sigma several times?

Harding: Most of these I have forgotten. I'm 76 and most were over 30, 40 and up to over 75 years ago!

Jacobsen: For the ESOTERIQ Society, it states, "*Christopher Harding* (Australia): 197 on SBIS-Oxford-Analysis-New-Zealand in 1976." What is the full name of the SBIS-Oxford-Analysis-New-Zealand, particularly the "SBIS" part?

Harding: Don't know.

Jacobsen: While, fundamentally, dispensing with ethical philosophy, social philosophy, economic philosophy, political philosophy, and metaphysics, even philosophy as "word juggling" (!), I see some common points. One is science, though "less and less" with its despoilment, meaning as a "PATTERN" made by each person individually, an emphasis on some of Freud, "QUANTUM PHYSICS" in terms of "truth" with its preservation of information (neither gained nor lost), and the bounded nature of nature (including humans) as "a condition of being defined." So, there *is* a *there* there. I have to ask, "What makes these conclusions more sound, at this time, to you than other possibilities?

Harding: Feynman once said no one understands the Quantum. And yet to further agree with his point "Quantum Superiority" has been proved for the D-Wave Orion Computer. I liken this to statements about the Aleph series in the mathematics of infinity theory.

Jacobsen: Any speculation as to why the information in Quantum Physics simply "IS"?

Harding: I once thought it through and concluded there was another stage beyond Quantum Physics. Simply IS would represent in turn a `single one' off any general group of abstractions.

Dr. Ronald K. Hoeflin

2024-08-22

Original publication in In-Sight 1, 2, and 3,

Dr. Ronald K. Hoeflin founded the Prometheus Society and the Mega Society, and created the Mega Test and the Titan Test. Hoeflin discusses: family geographic, cultural, linguistic, and religious background; depth of known family history; feelings about some distinguished family members in personal history; upbringing for him; discovery and nurturance of giftedness; noteworthy or pivotal moments in the midst of early life; early aptitude tests; inspiration for the Mega Society – its title, rarity, and purpose; inspiration for the Prometheus Society – its title, rarity, and purpose; inspiration for the Top One Percent Society – its title, rarity, and purpose; inspiration for the One-in-a-Thousand Society – its title, rarity, and purpose; inspiration for the Epimetheus Society – its title, rarity, and purpose; inspiration for the Omega Society – its title, rarity, and purpose; the developments of each society over time; communications of high-IQ societies, and harshest critiques of high-IQ societies; overall results of the intellectual community facilitated for the gifted; Prometheus Society and the Mega Society kept separate from the Lewis Terman Society, and Top One Percent Society, One-in-a-Thousand Society, Epimetheus Society, and Omega Society placed under the aegis of the “The Terman Society” or

"The Hoeflin Society"; disillusionment with high-IQ societies; notable failures of the high-IQ societies; changing norms of the Mega Test and the Titan Test; the hypothetical Holy Grail of psychometric measurements; other test creators seem reliable in their production of high-IQ tests and societies with serious and legitimate intent respected by Dr. Hoeflin: Kevin Langdon and Christopher Harding; societies societies helpful as sounding boards for the *Encyclopedia of Categories*; librarian work helpful in the development of a skill set necessary for independent psychometric work and general intelligence test creation; demerits of the societies in personal opinion and others' opinions; virtues and personalities as mostly innate or inborn, and dating and mating; publications from the societies attempted to be published at a periodic rate; faux and real genius; validity to Professor Robert Sternberg's Triarchic Theory of intelligence with practical intelligence, creative intelligence, and analytical intelligence; validity to Multiple Intelligences Theory of Professor Howard Gardner with musical-rhythmic, visual-spatial, verbal-linguistic, logical-mathematical, bodily-kinesthetic, interpersonal, intrapersonal, naturalistic, existential, and teaching-pedagogical intelligences; validity to general intelligence, or g, of the late Charles Spearman; the general opinion on the three main theories of intelligence; self-identification as a genius; personal opinions on the state of mainstream intelligence testing and alternative high-range intelligence testing; statistical rarity for apparent and, potentially, actual IQ scores of females who score at the extreme sigmas of 3, 4, and 5, or higher; reducing or eliminating social conflicts of interest in test creation; multiple test attempts; data on the Mega Test and the Titan Test; pseudonyms and test scores; and possible concerns of the test creators at the highest sigmas.

Scott Douglas Jacobsen: In due course of this personal and educational comprehensive interview, we will focus, in-depth, on the monumental life work of the (currently) 10-volume *The Encyclopedia of Categories* – a truly colossal intellectual endeavour. You founded some of the, if not the, most respected general intelligence tests in the history of non-mainstream general intelligence testing: The Mega Test and the Titan Test. Also, you founded the Mega Society in 1982. Another respected product of a distinguished and serious career in the creation of societies for community and dialogue between the profoundly and exceptionally gifted individuals of society. Before coverage of this in the interview, let's cover some of the family and personal background, I intend this as comprehensive while steering clear of disagreements or political controversies between societies, or clashes between individuals in the history of the high IQ societies – not my territory, not my feuds, not my business. Almost everything at the highest sigmas started with you [Ed. some integral founders in the higher-than-2-sigma range include Christopher Harding and Kevin Langdon], as far as I can tell, I want to cover this history and give it its due attention. What was family background, e.g., geography, culture, language, and religion or lack thereof?

Dr. Ronald K. Hoeflin: I recently wrote a 51-page autobiographical sketch for inclusion in my upcoming multi-volume treatise titled *The Encyclopedia of Categories,* a 10-volume version of which will probably be available for free as ten email attachments by January of 2020. I was aiming for a 13-volume version, but I don't think I can complete that length before the end of 2020. Given that my vision is way below 20/20, I liked the irony of publishing this final magnum opus of mine in the year 2020. I can always stretch it to 13 or more volumes in subsequent

editions. I will not quote what I say in that autobiographical sketch, although the information provided will be roughly the same. My mother's ancestors came from the British Isles (England, Scotland, and Ireland) mostly in the 1700s. My mother's father was a hellfire-and-brimstone Southern Methodist itinerant preacher in the state of Georgia. He's the only one of my four grandparents I never met. My mother brought me up as a Methodist, but I asked a lot of questions by my mid-teens and became a complete atheist by the age of 19, which I have remained ever since (I'm now 75). I gave my mother Bertrand Russell's essay "Why I Am Not a Christian" to read aloud to me so we could discuss it. It seemed to convince her to give up religion, which shows unusual flexibility of mind for a person in her 50s. She had previously read such books as *The Bible as History* and Schweitzer's *Quest of the Historical Jesus*, his doctoral dissertation in theology. My father's parents came to this country in the late 1890s, his mother from the Zurich region of Switzerland and his father from the Baden region of Germany. His father was a pattern maker, a sort of precision carpentry in which he made moulds for machine parts to be poured from molten metal in a foundry. My father became an electrical engineer, initially working on power lines in the state of Missouri, then becoming a mid-level executive for the main power company in St. Louis, Missouri, doing such things as preparing contracts with hospitals for emergency electrical power *generation* if the main city-wide power cut off. He had worked his way through college by playing the violin for dance bands, and as an adult he taught ballroom dancing in his own studio as a hobby. My mother was an opera singer. In my autobiography, I list the 17 operas she sang in during her career, usually with leading roles due to the excellence of her voice. My father initially spoke German up to the age of 2, but his parents decided they did not want their daughter doing so, so they started speaking English at home, so she never learned German. My father's mother became a devoted Christian Scientist and got her husband and two daughters to adopt this religion. My father became an atheist, and when he heard that my brother was thinking of becoming a Methodist minister sent him a copy of Thomas Paine's book *The Age of Reason*, which promotes Paine's deism, in which he accepted a deity and an afterlife but rejected the Bible as a guide, regarding the universe itself as God's true bible. My brother never read the book but I did, and I told my father I enjoyed the critique of the Bible but did not accept a God or afterlife, and my father said that these two beliefs could readily be discarded, but that Paine should be given credit for his advanced thinking in an era and country that so fiercely rejected atheism. My brother ultimately became a computer programmer for the pension system for employees of the state of California. My sister became a ballet dancer for the Metropolitan Opera in New York City. I list 25 operas she danced in in my autobiography. She went on to teach ballet at an upstate New York college, being honored one year as the college's most distinguished teacher.

Jacobsen: How far back is knowledge of the family history for you?

Hoeflin: I don't know much beyond what is stated above. My sister has more detailed records. One of my mother's grandfathers apparently owned over a hundred slaves in the South before the Civil War. My mother was occasionally treated badly in St. Louis due to her Southern accent, but she actually was very kindly toward black people and she once gave a black woman a ride in her car for a mile or so while I moved to the back seat. I do have memories of visits to my mother's mother in Atlanta, Georgia. She died before my third birthday, but my memories go back much

further than is normal with most people. I liked to swing on the swing in my mother's mother back yard with one of her chickens in my lap. She raised the chickens to sell their eggs, but evidently also killed them for dinner. I am even now very tender-hearted towards animals and would never kill a chicken or cow or what have you. But I still do eat meat out of habit, even though I regard it as not very ethical to do so. If I had a better income I'd arrange to eat just a vegetarian diet, mostly fruits and oatmeal. I loathe cooked green vegetables except in soups.

Jacobsen: Some harbour sentiments and feelings based on distinguished family members from centuries or decades ago. Those who died with great achievements or honourable lives in the sense of a well-lived life – whether prominent or not. Any individuals like this for you? Any sentiments or feelings for you?

Hoeflin: A genealogist traced my mother's ancestors to a close relative of a governor of Virginia. My mother said some of her relatives were distinguished doctors (M.D.s). I have a close friend who lives in Poland now, where she was raised, who is a great-great-great-great granddaughter of Catherine the Great (one of her great-grandmothers was a great-granddaughter of Catherine the Great). She shares a surprising number of characteristics that Catherine had despite the rather distant ancestry: a significant talent for learning languages, a love of art, an imperious attitude, and an embarrassing number of superstitions. I also dated a woman who was an out-of-wedlock daughter of Pablo Picasso, and there again there were striking similarities between the daughter and her father, even though she did not learn from her mother that he was her real father until 1988, some 15 year after his death in 1973. She started out as a virtuoso violinist, but by her 20s became a painter and had works of art in five different museums by the time she learned who her true father was. She also had facial features very much like Picasso's, even though she was raised in a German family. I am proud that my mother and sister were so gifted in their respective arts (singing and ballet). When I drew up a list of my favourite classical musical pieces for my autobiography, I looked at YouTube to see the actual performances, and it struck me what a lot of amazingly talented people could perform these magnificent pieces of music, and I regret how limited I am in my talents. I can't even drive a car due to my poor eyesight! It is chiefly or only in these incredible aptitude test scores that I seem to shine way beyond the norm. I read when I was in high school that the average high-school graduate could read 350 words per minute, so I tested myself, and I found that on a few pages of a very easy sci-fi novel I could read only 189 words per minute at top speed, which works out to just 54% as fast as the average high-school graduate. Yet on timed aptitude tests as a high-school sophomore, I reached the 99th percentile in verbal, spatial, and numerical aptitude despite this huge speed deficit. And on the verbal aptitude section of the Graduate Record Exam I reached the top one percent compared to college seniors trying to get into graduate school, an incredible achievement given my dreadful reading speed. As I mention in my autobiographical sketch, if I had to read aloud, even as an adult I read so haltingly that one would assume that I am mentally retarded if one did not know that the cause is poor eyesight, not poor mental ability.

Jacobsen: What was upbringing like for you?

Hoeflin: My parents were divorced when I was 5 and my mother went through hours-long hysterical tantrums every 2 or 3 weeks throughout my childhood, which were emotionally

traumatic and nightmarish. My father had an affable and suave external demeanour but was very selfish and cruel underneath the smooth facade. My brother pushed me downstairs when I was 3 and I stuck my forehead on the concrete at the bottom, causing a gash that had to be clamped shut by a doctor. It was discovered that I had a detached retina when I was 7 (because I could not read the small print in the back of the second-grade reader that the teacher called on me to read), and I spent my 8th birthday in the hospital for an eye operation, for which my father refused to pay since he did not believe in modern medicine, just healthy living as the cure for everything. So even though he was an engineer, my mother had a more solid grasp of physical reality than he did, as I mentioned to her once. I flunked out of my first and third colleges due in large measure to my visual problems, but I eventually received two bachelor's degrees, two master's degrees, and a doctorate after going through a total of eight colleges and universities. So all in all my childhood was rocky and unpleasant. As an adult, I took the personality test in the book *Personality Self-Portrait* and my most striking score was on a trait called "sensitivity," on which I got a perfect score of 100%. On the twelve other traits, I scored no higher than 56% on any of them. I never tried sexual relations until the age of 31, and I found that I could never reach a climax through standard intercourse. I had a nervous breakdown after trying group psychotherapy for a few sessions when the group's criticism of the therapist after he left the room reminded me of my mother's criticisms of my father, crying for 12 hours straight. When I mentioned this at the next therapy session, one of the other people in the group came up to me afterward and told me he thought I was feeling sorry for myself, despite the fact that my report to the group was very unemotional and matter-of fact, not dramatic. I accordingly gave up group therapy after that session. On the personality test, on the trait called "dramatic", I actually scored 0%, probably because pretending to be unemotional discourages needling from sadistic people who love to goad a highly sensitive person like me.

Jacobsen: When was giftedness discovered for you? Was this encouraged, supported, and nurtured, or not, by the community, friends, school(s), and family?

Hoeflin: At the age of 2 my mother's mother picked me up when I was running to her back yard upon arriving in Atlanta to grab one of her chickens to swing with it on my lap. At first I ignored her, but then I surmised that she wanted to ask me a question, so I looked at her face, waiting for her question, which never came. Maybe she didn't realize that my command of the language had improved since my previous visit. She eventually tapped me on the head and told my mother "You don't have to worry about this one, he's got plenty upstairs." My mother told me this story several times over the years, and I finally put two and two together and told my mother I recalled the incident, which shocked her considering how young I had been. I told her that her mother had probably been impressed by my long attention span. My mother then thought that the incident was not as important and mysterious as she has thought, but actually a long attention span at such a young age is probably a good sign of high intelligence. It was not until I was in the fifth grade that I was given aptitude tests and the teacher suddenly gave me eighth-grade reading books and sixth-grade math books. This was in a so-called "sight conservation class" for the visually impaired that I attended in grades 3 through 5. The teacher taught students in grades 1 through 8 in a single classroom because very poor vision is fairly rare even in a city as large as St. Louis, at that time the tenth-largest city in the United States. That gave me plenty of time to explore my

own interests, such as geography using the world maps they had on an easel. In grade 8, back in a regular classroom, we were given another set of aptitude tests, and the teacher mentioned to the class that I had achieved a perfect score on a test of reading comprehension, meaning I was already reading at college level. The teacher gave us extra time on the test so I would have time to finish the test. A problem toward the end of the test clued me in on how to solve a problem that had stumped me earlier in the test, so I went back and corrected that previous answer. Then there were those three 99th percentile scores as a high-school sophomore that I've already mentioned. When I learned that my reading speed was so slow compared to others, I realized that my true aptitudes (minus the visual handicap) must be well within the top one percent on each of the three tests.

Jacobsen: Any noteworthy or pivotal moments in the midst of early life in school, in public, with friends, or with family?

Hoeflin: In the seventh grade I suddenly started creating crossword puzzles and mazes, a harbinger of my later creation of the two tests that appeared in *Omni* magazine in April 1985 and in April 1990. I also collected lists of fundamental things such as independent countries of the world, the Western Roman emperors, the chemical elements, the planets and their moons, etc., in keeping with my much earlier childhood ambition to know everything. If you can't know everything, then at least know the basic concepts for important subjects like geography, history, chemistry, astronomy, etc. These lists were a harbinger of my current multi-volume treatise on categories.

Jacobsen: Were there early aptitude tests of ability for you? What were the scores and sub-test scores if any? Potentially, this is connected to an earlier question.

Hoeflin: The only other test I should mention is the Concept Mastery Test. Lewis Terman collected a group of 1,528 California school children in grades 1 through 12 with IQs in the 135 to 200 range. To test their abilities as adults he and his colleagues constructed two 190-problem tests covering mostly vocabulary and general knowledge, which are easy problems to construct but are known to correlate well with general intelligence, the first test (Form A) administered to his group in 1939-1940 and the second one (Form B, latter called Form T) in 1950-52. About 954 members of his group tried the first one and I think 1,024 tried the second test. But Terman made the second test much easier than the first in order to make it easier to compare his group to much less intelligent groups such as Air Force captains. So the Mensa (98th percentile) cut-off would be a raw score of about 78 out of 190 on the first test and about 125 out of 190 on the second. I was editor for the Triple Nine Society (minimum requirement: 99.9 percentile) for a few years starting in 1979, and some members sent me copies of the two CMT tests so I could test TNS members. Since the CMT tests were untimed, I was not handicapped by the speed factor. Compared to Terman's gifted group I reached the top one percent on both tests. According to Terman's scaling of Form A, my raw score of 162.5 would be equivalent to an IQ of 169.4 (assuming a mean of 100 and a standard deviation of 16 IQ points), where an IQ of 168.3 would be equivalent to the 99.999 percentile or one-in-100,000 in rarity. By comparing adult CMT IQs with childhood Stanford-Binet IQs for Terman's group, I calculated that my adult 169.4 IQ would be equivalent to a childhood IQ of 192. The one-in-a-million level on the two tests (the

99.9999 percentile) would be about 176 IQ on the CMT and 204 IQ on the Stanford-Binet, respectively.

The *Guinness Book of World Records* abandoned its "Highest IQ" entry in 1989 because the new editor thought (correctly) that it is impossible to compare people's IQs successfully at world-record level. The highest childhood IQ I know of was that of Alicia Witt, who had a mental age of 20 at the age of 3. Even if she had been 3 years 11 months old, this would still amount to an IQ of over 500! At the age of 7, she played the super-genius sister of the hero in the 1984 movie *Dune.* On a normal (Gaussian) curve such an IQ would be impossible since an IQ of 201 or so would be equivalent to a rarity of about one-in-7-billion, the current population of the Earth. But it is well known to psychometricians that childhood IQs using the traditional method of mental age divided by chronological age fail to conform to the normal curve at high IQ levels. The Stanford-Binet hid this embarrassing fact in its score interpretation booklet (which I found a copy of in the main library of the New York Public Library) by not awarding any IQs above 169, leaving the space for higher IQs blank! The CMT avoids the embarrassment of awarding IQs of 500 or more by having a maximum possible IQ on Form A (the harder of the two CMTs) of 181. Leta Speyer and Marilyn vos Savant, both of whom I had dated for a time, had been listed in the *Guinness Book of World Records* as having world-record IQs of 196 and of 228, respectively, Marilyn having displaced Leta in the 1986 edition. Leta felt that the 228 IQ of Marilyn was fake, but I was aware that these childhood scores could go well beyond 200 IQ because they fail to conform to the normal curve that Francis Galton had hypothesized as the shape of the intelligence curve in his seminal book *Hereditary Genius* (first edition 1869, second edition 1892). I was unable to contact Alicia Witt to see if she would be interested in joining the Mega Society. I should note that the three key founders of the ultra-high-IQ societies (99.9 percentile or above) were Chris Harding, Kevin Langdon, and myself. Harding founded his first such society in 1974, Langdon in 1978, and myself in 1982. Mensa, the granddaddy of all high-IQ societies with a 98th percentile minimum requirement, was founded in 1945 or 1946 by Roland Berrill and L. L Ware, and Intertel, with a 99th percentile minimum requirement, was founded in 1966 or 1967 by Ralph Haines. I don't care to quibble about the precise dates that Mensa and Intertel were founded, so I have given two adjacent dates for each. In its article "High IQ Societies" Wikipedia lists just 5 main high-IQ societies: Mensa, Intertel, the Triple Nine Society, the Prometheus Society, and the Mega Society (minimum percentile requirements: 98, 99, 99.9, 99.997, and 99.9999, respectively; or one-in 50, one-in-100, one-in-1,000, one-in-30,000, and one-in-1,000,000; dates founded: roughly 1945, 1966, 1979, 1982, and 1982; founders: Berrill and Ware, Haines, Kevin Langdon, Ronald K. Hoeflin, and Ronald K. Hoeflin, respectively.

Jacobsen: Perhaps, we can run down the timeline of the six societies in this part with some subsequent questions: Prometheus Society (1982), Mega Society (1982), Top One Percent Society (1989), One-in-a-Thousand Society (1992), Epimetheus Society (2006), and Omega Society (2006). What was the inspiration for the Mega Society – its title, rarity, and purpose?

Hoeflin: Kevin Langdon had a list of 600 or so people who had qualified for his Four Sigma Society from the 25,000 *Omni* readers who tried his LAIT (Langdon Adult Intelligence Test) that appeared in *Omni* in 1979. Four Sigma was given a cut-off of four standard deviations above the mean, which on a normal curve would be about one-in-30,000 in rarity or the 99.997 percentile.

So approximately one-thirtieth of them should have been qualified for a one-in-a-million society. I suggested to him that he might ask the top 20 scorers if they'd like to form the nucleus of a one-in-a-million society, but he evidently thought this cut-off was too high to be practical. So when he let his Four Sigma Society languish, I decided to start Prometheus as a replacement for it, with the Mega Society as a follow-through on my suggestion to him about starting a one-in-a-million society, where "mega" means, of course, "million," indicating how many people each member would be expected to exceed in intelligence. With slightly over 7 billion people, there would be a pool of about 7,000 potential Mega Society members, or slightly less if we exclude young children. I knew of a statistical method by which several very high scores from several tests could be combined to equal a one-in-a-million standard, as if the several tests constituted a single gigantic test. So I accepted members using this statistical method until my Mega Test appeared in *Omni* in April 1985. I put the cut-off at a raw score of 42 out of 48 initially, but then increased this to 43 after getting a larger sample. The test was eventually withdrawn from official use for admission to the Mega Society because some psychiatrist maliciously published a lot of answers online that others could search out and copy. At this time my other test, the Titan Test, is the only one that the Mega Society will accept, again at a raw score of 43 out of 48.

Jacobsen: What was the inspiration for the Prometheus Society – its title, rarity, and purpose?

Hoeflin: The Prometheus Society, as mentioned above, was intended as a replacement for the Four Sigma Society, which Langdon had allowed to languish. Prometheus was a figure in Greek mythology who was punished by the gods for giving fire to humans. I told Kevin, half in jest, that I was stealing his idea for the Four Sigma Society from him like Prometheus stealing fire from the gods! On my Mega and Titan Test, the qualifying score for Prometheus is a raw score of 36 out of 48, roughly equivalent to a rarity of one-in-30,000 or the 99.997 percentile, the same as Four Sigma's cut-off, i.e., a minimum qualifying score.

Jacobsen: What was the inspiration for the Top One Percent Society – its title, rarity, and purpose?

Hoeflin: I wanted to make a living publishing journals for high-IQ societies. I initially was able to do so as the editor for the Triple Nine Society, for which I was paid just $1 per month per member for each monthly journal I put out. When I started as editor in late 1979, there were only about 50 members, but once Kevin's test appeared in *Omni* the number of members swelled to about 750. With $750 per month, I could put out a journal and still have enough left over to live on, since my monthly rent was just $75 thanks to New York City's rent laws. When Kevin heard that I was able to do this, he was not amused, since he thought the editorship should be an unpaid position. So I started the Top One Percent Society from people who had taken my Mega Test in *Omni* in April 1985 and my Titan Test in April 1990, thus removing myself from any disputes with Kevin or other members of the Triple Nine Society. I liked being self-employed rather than work as a librarian, which had been my profession from 1969 to 1985, because difficulties with higher-ups in the library field could crop up if there were personality conflicts.

Jacobsen: What was the inspiration for the One-in-a-Thousand Society – its title, rarity, and purpose?

Hoeflin: I started the One-in-a-Thousand Society when income from my Top One Percent Society started to seem insufficient, even when I put out two journals per month rather than one for the Top One Percent Society. The third journal per month was a bit more hectic, but within my capacity.

Jacobsen: What was the inspiration for the Epimetheus Society – its title, rarity, and purpose?

Hoeflin: In Greek mythology, Epimetheus was a brother to Prometheus. I'd let the Prometheus and Mega societies fall into the control of other people, so I decided to create new societies at their same cut-offs but with different names and under my control. I don't recall the motivation for founding Epimetheus, since starting in 1997 I qualified for Social Security Disability payments due to my poor vision and low income, and that completely solved all my financial worries, even when my rent gradually crept up from $75 to $150 from 1997 to around 2003. It is now permanently frozen at $150 a month due to an agreement with an earlier landlord, who wanted the City to give him permission to install luxury apartments where I live, for which he could charge $2,000 to $4,000 a month due to the proximity to Times Square, which is just ten minutes' walk away. I think that the Prometheus Society was restricting the tests it accepted to just a very small number of traditional supervised IQ tests, excluding unsupervised amateur-designed tests like mine. I wanted my tests to still serve a practical purpose at the Prometheus and Mega cut-offs.

Jacobsen: What was the inspiration for the Omega Society – its title, rarity, and purpose?

Hoeflin: Chris Harding of Australia was forever founding new high-IQ societies with new names but whose existence was largely known only to him and the people he awarded memberships to. He founded an Omega Society at the one-in-3,000,000 cut-off, but I assumed after several years of hearing nothing about it that it must be defunct, so I decided to call my new one-in-a-million society the Omega Society, since "Omega" seemed a nice twin word for "Mega" just as "Epimetheus" served as a twin word for "Prometheus." Chris wrote to me about this appropriation of his society's name and I explained my reason for adopting it. He offered no further complaint about it.

Jacobsen: What were the developments of each society over time?

Hoeflin: I decided to devote my full-time attention to a massive multi-volume opus titled "The Encyclopedia of Categories," of which I'd published a couple of one-volume versions in 2004 and 2005. When I noticed that Samuel Johnson's great unabridged dictionary of 1755 could now be bought for just $9.99 from Kindle, the computer-readable format that avoids paper printing, I decided I could make an affordable multi-volume treatment of my "Encyclopedia of Categories." I'd also discovered that quotations from collections of quotations could be analyzed in terms of my theory of categories, giving me a virtually inexhaustible source of examples considering how many quotation books there are out there. So I sold the four societies that were still under my control to Hernan Chang, an M.D. physician living in Jacksonville, Florida, as well as all of my IQ tests. Although, he lets me score the latter for him and collect the fee, since he is too busy to handle that. I began my multi-volume opus in late 2013 and believe I can complete a 10-volume version by the end of this year, 2019. I was initially aiming at a 13-volume version, in harmony

with the number of basic categial niches I employ, but it would take until early 2021 to complete the extra 3 volumes, so I'll publish a 10-volume version in January of 2020. The year 2020 as a publication date appealed to me because of its irony, given that my visual acuity falls far short of 20/20, and the year 2020 rolls around only once in eternity, if we stick to the same calendar. I could still put out more volumes in later editions if I felt so inclined, but I let readers voice an opinion on the optimum number of volumes.

Jacobsen: What was the intellectual productivity and community of the societies based on self-reports of members? What have been the harshest critiques of high IQ societies from non-members, whether qualifying or not?

Hoeflin: I think the focus of the higher-IQ societies has been on communication with other members through the societies' journals. I never tried to keep track of the members' "intellectual productivity." As for harsh critiques of the high-IQ societies, the only thing that comes to mind is *Esquire* magazine's November 1999 so-called "Genius" issue. It focused on four high-IQ-society members, including myself. I never read the issue except for the page about myself, and it took me two weeks to get up enough nerve to read even that page. I was told by others that the entire issue was basically a put-down of high-IQ societies and their members, although people said the treatment of me was the mildest of the four. I did notice that they wanted a photo of me that looked unattractive, me using a magnifying glass to read. I suggested a more heroic picture, such as me with one of my cats, but they kept taking pictures of me peering through that magnifying glass in a rather unflattering pose, with zero interest in alternative poses. Kevin Langdon was sarcastic about our willingness to expose ourselves to such unflattering treatment. (He was not among the four that they covered in that issue.)

Jacobsen: What have been the overall results of the intended goals of the provision of an intellectual community of like-gifted people who, in theory, may associate more easily with one another? I remain aware of skepticism around this idea, which may exist in the realm of the naive.

Hoeflin: I had found that I could not interact with members of Mensa, who generally treated me as a nonentity. I was also very shy and unable to put myself forward socially in Mensa groups. At the higher-IQ levels, however, I had the prominent role of editor and even founder, which made it possible for others to approach me and break through that shyness of mine. So I did manage to meet and interact with quite a few people by virtue of my participation in the high-IQ societies, although the ultimate outcome seems to be that I will probably end my life in total isolation from personal friends except a few people who reach out to me by phone or email, as in the present question-and-answer email format. As for other people, they will have to tell you their own stories, since people are quite diverse, even at very high IQ levels.

Jacobsen: Why were the Prometheus Society and the Mega Society kept separate from the Lewis Terman Society? Why were the Top One Percent Society, One-in-a-Thousand Society, Epimetheus Society, and Omega Society placed under the aegis of the Lewis Terman Society? Also, what is the Lewis Terman Society?

Hoeflin: I think Hernan Chang adopted the name "The Hoeflin Society" in preference to "The Terman Society" as an umbrella term for the four societies he purchased from me.

Jacobsen: What have been the merits of the societies in personal opinion and others' opinions?

Hoeflin: Speaking personally, I have lost almost all interest in the high-IQ societies these days, although I am still a nominal, non-participatory member of several of them. One group I joined recently as a passive member named the "Hall of Sophia" unexpectedly offered to publish my multi-volume book in any format I like for free. The founder had taken my Mega or Titan test earlier this year (February 2019) and did quite well on it, and was sufficiently impressed to classify me as one of the 3 most distinguished members of his (so far) 28-member society. I was going to send out my book for free as email attachments fo people listed in the *Directory of American Philosophers* as well as to any high-IQ-society members who might be interested. So for me, the one remaining merit of the high-IQ societies would be to have a potential audience for my philosophical opus.

Jacobsen: When did you begin to lose interest or become disillusioned, in part, in high-IQ societies? My assumption: not simply an instantaneous decision in 2019.

Hoeflin: Editing high-IQ-society journals from 1979 onwards for many years, at first as a hobby and then as a livelihood, kept me interested in the high-IQ societies. I gave up the editing completely around 2009. Thirty years is plenty of time to become jaded. Getting Social Security Disability payments in 1997 removed any financial incentive for publishing journals. Over the years I'd travelled to such destinations as California and Texas and Illinois for high-IQ-society meetings, not to mention meetings here in New York City, when I had sufficient surplus income, but all things peter out eventually.

Jacobsen: What have been the notable failures of the high-IQ societies?

Hoeflin: There was actually talk of a commune-like community for high-IQ people, but after I saw how imperious some high-IQ leaders like Kevin Langdon were, this would be like joining Jim Jones for a trip to Guyana–insane! That's hyperbole, of course. Langdon actually ridiculed the followers of Jim Jones for their stupidity in following such a homicidal and suicidal leader, not to mention his idiotic ideas. Langdon advocates a libertarian philosophy, but in person he is very controlling. I guess we just have to muddle through on our own, especially if we have some unique gift that we have to cultivate privately, not communally. Langdon often ridiculed my early attempts to develop a theory of categories, but I'm very confident in the theory now that I have worked at it for so long. Human beings tend to organize their thoughts along the same systematic lines, just like birds instinctively know how to build nests, spiders to build webs, and bees to build honeycombs. My analyses are so new and startling that I'm sure they will eventually attract attention. If I'd been an epigone of Langdon, I'd never have managed to develop my theory to its present marvellous stage.

Jacobsen: With the Flynn Effect, does this change the norms of the Mega Test and the Titan Test used for admissions purposes in some societies at the highest ranges?

Hoeflin: A lot of people suddenly started qualifying for the Mega Society, perhaps from copying online sources or perhaps from the test suddenly coming to the attention of a lot of very smart

people. So initially higher scores on that test were required and then the test was abandoned entirely as an admission test for the Mega Society. Terman found that his subjects achieved gradually higher IQ scores on his verbal tests the older they got. One theory is that as people gradually accumulate a larger vocabulary and general knowledge (crystallized intelligence) their fluid intelligence, especially on math-type tests, gradually declines, so that if one relies on both types of intelligence, then your intelligence would remain relatively stable until extreme old age. There has been no spurt in extremely high scores on the Titan Test, however.

Jacobsen: What would be the Holy Grail of psychometric measurements, e.g., a non-verbal/culture fair 5-sigma or 6-sigma test?

Hoeflin: The main problem with extremely difficult tests is that few people would be willing to attempt them, so norming them would be impossible. I was astonished that the people who manage the SAT have actually made the math portion of that test so easy that even a perfect score is something like the 91st percentile. Why they would do such an idiotic thing I have no idea. Terman did the same thing with his second Concept Mastery Test, so that a Mensa-level performance on that test would be a raw score of 125 out of 190, whereas a Mensa-level performance on the first CMT was 78 out of 190. Twenty members of his gifted group had raw scores of 180 to 190 on the second CMT whereas no member of his group had a raw score higher than 172 out of 190 on the first CMT. His reason was to be able to compare his gifted group with more average groups such as Air Force captains, who scored only 60 out of 190 on the second test, less than half as high as Mensa members. A lot of amateur-designed intelligence tests have such obscure and difficult problems that I am totally unable to say if those tests have any sense to them or not. Perhaps games like Go and Chess are the only ways to actually compare the brightest people at world-record levels. But such tests yield to ever-more-careful analysis by the competitors, so that one is competing in the realm of crystallized intelligence (such as knowledge of chess openings) rather than just fluid intelligence. Even the brightest people have specialized mental talents that help them with some tests but not with others, like people who compete in the Olympic Decathlon, where some competitors will do better in some events and others in other events, the winner being the one with the best aggregate score. General intelligence means that even diverse tests like verbal, spatial, and numerical ones do have some positive intercorrelation with each other–they are not entirely independent of each other. The best tests select problems that correlate best with overall scores. But few if any of the amateur-designed tests have been subjected to careful statistical analysis. Some people did subject my Titan Test to such statistical analysis and found that it had surprisingly good correlations with standard intelligence tests, despite its lack of supervision or time limit.

Jacobsen: Other than some of the work mentioned. What other test creators seem reliable in their production of high-IQ tests and societies with serious and legitimate intent? Those who you respect. You have the historical view here – in-depth in information and in time. I don't.

Hoeflin: I think Kevin Langdon's tests are very well made and intelligent, but he tends to focus on math-type problems. Christopher Harding, by contrast, focuses on verbal problems and does poorly in math-type problems. For international comparisons across languages, I guess one would have to use only math-type problems, as I did in my Hoeflin Power Test, which collected

the best math-type problems from the three previous tests (Mega, Titan, and Ultra). But English is virtually a universal language these days, so perhaps verbal tests that focus on English or perhaps on Indo-European roots could be used for international tests, except that Indo-European languages constitute only 46% of all languages, by population. I think Chinese will have difficulty becoming culturally dominant internationally because the Chinese language is too difficult and obscure for non-Chinese to mess with.

Jacobsen: Were the societies helpful as sounding boards for the *Encyclopedia of Categories*?

Hoeflin: I used high-IQ-society members as guinea pigs to develop my intelligence tests, but my work on categories I have pursued entirely independently, except for the precursors I rely on, notably the philosopher Stephen C. Pepper (1891-1972), who taught at the University of California at Berkeley from 1919 to 1958. Oddly enough, in his final book titled *Concept and Quality* (1967) he used as a central organizing principle for his metaphysics what he called "the purposive act," of which he said on page 17: "It is the act associated with intelligence"!!! I simply elaborated this concept from 1982 when I first read *Concept and Quality* onward, elaborating it into a set of thirteen categories by means of which virtually any complete human thought or action, as in a quotation, can be organized. In my introductory chapter, which currently traces the development of my theory from William James last book, *A Pluralistic Universe*, to the present, I now plan to trace the thirteen categories not just to the Greeks and Hebrews but back to animal life and ultimately back to the Big Bang, breaking the stages of its development into 25 discrete ones including my own contributions toward the end. I may begin with Steven Weinberg's book *The First Three Minutes* and end with Paul Davies kindred book, *The Last Three Minutes,* if I can manage to extract convincing 13-category examples from each of these books.

Jacobsen: How was librarian work helpful in the development of a skill set necessary for independent psychometric work and general intelligence test creation?

Hoeflin: It was mostly helpful to me because I could work part-time during the last ten years of my 15 or 16 years as a librarian, which gave me the leisure for independent hobbies, thought, and research.

Jacobsen: What have been the demerits of the societies in personal opinion and others' opinions?

Hoeflin: There tends to be a lot of arrogance to be found among members of the high-IQ societies, so charm is typically not one of their leading virtues. They generally assume that virtually everyone they speak to is stupider than they are.

Jacobsen: How can members be more humble, show more humility? Also, what are their leading virtues?

Hoeflin: I think personalities are largely inborn and can't be changed much. Perhaps there should be sister societies, analogous to college sororities, for women who have an interest in socializing with high-IQ guys for purposes of dating and mating. In the ultra-high-IQ societies, women constitute only about 6% of the total membership. (Parenthetically, if you look at the

Wikipedia list of 100 oldest living people, one usually finds about 6 men and 94 women.) In Mensa, the percentage of women typically ranges from 31% to 38%.

Jacobsen: How many publications come from these societies? What are the names of the publications and the editors in their history? What ones have been the most voluminous in their output – the specific journal? Why that journal?

Hoeflin: Each society generally has a journal that it tries to publish on a regular basis. Kevin Langdon puts out *Noesis*, the journal for the Mega Society, about twice per year. I also get journals from Prometheus and Triple Nine and Mensa. The four societies Hernan Chang operates all function entirely online, and I have never seen any of their communications. Even the journals I get I only glance at, never read all the way through. Due to my very slow reading speed, I tend to focus my reading on books that seem worthwhile from which to collect examples for my "Encyclopedia of Categories."

Jacobsen: Before delving into the theories, so a surface analysis, what defines a faux genius? What defines a real genius to you? Or, perhaps, what different definitions sufficiently describe a fake and a true genius for non-experts or a lay member of the general public – to set the groundwork for Part Three?

Hoeflin: I would say that genius requires high general intelligence combined with high creativity. How high? In his book *Hereditary Genius,* Francis Galton put the lowest grade of genius at a rarity of one in 4,000 and the highest grade at a rarity of one in a million. Scientists love to quantify in order to give their subject at least the appearance of precision. One in 4,000 would ensure one's being noticed in a small city, while one in a million would ensure one's being noticed in an entire nation of moderate size.

Jacobsen: By your estimation or analysis, any validity to Professor Robert Sternberg's Triarchic Theory of intelligence with practical intelligence, creative intelligence, and analytical intelligence?

Hoeflin: I like Sternberg's attempt at analyzing intelligence, but clearly just three factors seems a bit skimpy for a really robust theory.

Jacobsen: Any validity to Multiple Intelligences Theory of Professor Howard Gardner with musical-rhythmic, visual-spatial, verbal-linguistic, logical-mathematical, bodily-kinesthetic, interpersonal, intrapersonal, naturalistic, existential, and teaching-pedagogical intelligences?

Hoeflin: Here we have a more robust set of factors, but Gardner fails to show how his factors cohere within a single theory.

Jacobsen: Any validity to general intelligence, or g, of the late Charles Spearman?

Hoeflin: General intelligence was based on the fact that apparently quite diverse forms of intelligence such as verbal, spatial, and numerical have positive correlations between each pair of factors, presumably based on some underlying general intelligence.

Jacobsen: Amongst the community of experts, what is the general opinion on the three main theories of intelligence listed before? What one holds the most weight? Why that one?

Hoeflin: These are three theories in search of an overarching theory of intelligence. My guess is that the so-called "experts" lack the intelligence so far to create a really satisfactory theory of intelligence, perhaps analogous to the problem with finding a coherent theory of superstrings.

Jacobsen: Do you identify as a genius? If so, why, and in what ways? If not, why not?

Hoeflin: I think my theory of categories shows genuine genius. It even amazes me, as if I were just a spectator as the theory does its work almost independently of my efforts.

Jacobsen: Any personal opinions on the state of mainstream intelligence testing and alternative high-range intelligence testing now?

Hoeflin: I'm not up on the current state of intelligence testing. I do feel that it has focused way too much on the average range of intelligence, say from 50 to 150 IQ, i.e., from the bottom one-tenth of one percent to the top one-tenth of one percent. Testing students in this range is where the money is in academia. It's like music: all the money to be made is in creating pop music, which is typically of mediocre quality. Background music for movies is probably as close as music comes these days to being of high quality, presumably because there is money to be made from the movie studios in such music. I saw a movie recently called "Hangover Square," which came out in 1945. The title is unappealing and the movie itself is a totally unsuspenseful melodrama about a homicidal maniac whose identity is revealed right from the start. The one amazing thing about the movie was that the composer, Bernard Herman, composed an entire piano concerto for the maniac to purportedly compose and perform, with appropriate homicidal traits in the music to reflect the deranged soul of the leading character, the maniac. One rarely sees such brilliant musical talent thrown at such a horrible film. So I guess genius can throw itself into things even when the audience it is aimed at is of extremely mediocre quality. Maybe intelligence tests, even when they are aimed at mediocre students, can show glints of genius. The fact that I could attain the 99th percentile on tests aimed at average high-school students despite my slow reading due to visual impairment suggests that some psychometrician (or group of psychometricians) must have been throwing their creativity and intelligence into their work in an inspired way that smacks of true genius!

Jacobsen: Do the statistical rarities at the extreme sigmas have higher variance between males and females? If so, why? If not, why not? Also, if so, how is this reflected in subtests rather than simple composite scores?

Hoeflin: By "variance between males and females," I presume you are alluding to the fact that there tend to be more men at very high scores than women. This is especially obvious in spatial problems, as well as kindred math problems, presumably due to men running around hunting wild game in spatially complex situations while women sat by the fireside cooking whatever meat the men managed to procure. But it is also true that men outperform women on verbal tests. On the second Concept Mastery Test, a totally verbal test, of the 20 members of Terman's gifted group who scored from 180 to 190, the ceiling to the test, 16 were men but only 4 were women. This is a puzzling phenomenon, given women's propensity for verbalizing. Perhaps chasing game involves verbal communication, too, so that nature rewards the better verbalizers among

men in life-or-death situations. Warfare as well as hunting for game probably has a significant role in weeding out the unfit verbalizers among men.

Jacobsen: Following from the last question, if so, what does this imply for the statistical rarity for apparent and, potentially, actual IQ scores of females who score at the extreme sigmas of 3, 4, and 5, or higher?

Hoeflin: It obviously would be possible to breed women eugenically to increase the percentage of them with very high IQ scores. Even now, there are more women graduating from law school than men in the United States, which suggests no deficit in verbal intelligence at the high end of the scale. Although, there may be other reasons why men of high verbal intelligence avoid law as a career compared to women. Maybe, they are drawn away by other lucrative careers, such as business or medicine.

Jacobsen: In the administration of alternative tests for the higher ranges of general intelligence, individuals may know the test creator, even on intimate terms as a close colleague and friend. They may take the test a second time, a third time, a fourth time, or more. The sample size of the test may be very small. There may be financial conflicts of interest for the test creator or test taker. There may be various manipulations to cheat on the test. There may be pseudonyms used for the test to appear as if a first attempt at the alternative test. There are other concerns. How do you reduce or eliminate social conflicts of interest?

Hoeflin: Some people have used pseudonyms to take my tests when they were afraid I would not give them a chance to try the test a second or third time. There is not much incentive to score very high on these tests, except perhaps the prestige of joining a very high-IQ society. People cheat on standardized college admission tests, as we know from news reports, by getting other people to take the tests for them, for example. Considering how expensive colleges have become these days, my guess is that they will go the way of the dodo bird eventually, and people will get their education through computers rather than spending a fortune in a college. One guy cheated on my Mega Test by getting members of a think tank in the Cambridge, Massachusetts area to help him. He was pleased that I gave him a perfect score of 48 out of 48. He admitted cheating to Marilyn vos Savant, who informed me, so I disqualified his score. This was before my Mega Test appeared in *Omni.* Why he wanted credit for a perfect score that he did not deserve is beyond my understanding. I'd be more proud of a slightly lower score that I had actually earned. Another person has kept trying my tests, despite a fairly high scoring fee of $50 per attempt. I finally told him to stop taking the tests. His scores were not improving, so his persistence seemed bizarre.

Jacobsen: The highest score on the Mega Test on the first attempt by a single individual with a single name rather than a single individual with multiple names was Marilyn vos Savant at 46 out of 48. Similarly, with other test creators, and other tests, there were several attempts at the same test by others. Do the multiple test attempts and then the highest of those attempts asserted as the score for the test taker present an issue across the higher sigma ranges and societies?

Hoeflin: Some European guy did achieve a perfect score on the Mega Test eventually, about 20 years after the test first came out in 1985. The test is no longer used by any high-IQ societies that I know of due to the posting of mostly correct answers online by a malicious psychiatrist. He

probably needed to see a psychiatrist to figure out what snapped in his poor head to do such a thing. I guess it's a profession that attracts people with psychological problems that they are trying to understand and perhaps solve.

Jacobsen: What were the final sample sizes of the Mega Test and the Titan Test at the height of their prominence? How do these compare to other tests? What would be a reasonable sample size to tap into 4-sigma and higher ranges of intelligence with low margins of error and decent accuracy?

Hoeflin: A bit over 4,000 people tried the Mega Test within a couple of years of its appearance and about 500 people tried the Titan Test within a similar time period. Langdon's LAIT test is said to have had 25,000 participants. His test was multiple choice, whereas mine were not. A multiple-choice test is easier to guess on than a non-multiple-choice test. My tests were normed by looking at the previous test scores that participants reported and then trying to create a distribution curve for my tests what would jibe with the distribution on previously-taken tests. So I did not need to test a million or more people to norm my tests up to fairly high levels of ability.

Jacobsen: What are the ways in which test-takers try to cheat on tests? I mean the full gamut. I intend this as a means by which prospective test takers and society creators can arm themselves and protect themselves from cheaters, charlatans, and frauds, or worse. Same for the general public in guarding against them, whenever someone might read this.

Hoeflin: If people's wrong answers are too often identical with one another and out of sync with typical wrong answers, that is a clue that they are copying from one another or from some common source.

Jacobsen: Why do test takers use pseudonyms? How common is this practice among these types of test-takers? It seems as if a brazen and blatant attempt to take a test twice, or more, and then claim oneself as smart as the higher score rather than the composite of two, or more, scores, or even simply the lower score of the two, or more, if the scores are not identical.

Hoeflin: I know of a group of 5 M.I.T. students who collaborated and gave themselves the collective name of Tetazoo. There was also a professor at Caltech who tried the test but did not want his score publicized so he used the pseudonym Ron Lee. In both cases, the score just barely hit the one-in-a-million mark of 43 right out of 48. One person scored 42 right and wanted to try again so he used a pseudonym and managed to reach 47 right out of 48 on his second attempt.

Jacobsen: What have been and continue to be concerns for test creators at the highest sigmas such as yourself or others, whether active or retired? This is more of a timeline into the present question of the other suite of concerns.

Hoeflin: I do not know what are the main concerns of test designers, past or present, other than myself. I was fortunate to have Triple Nine members as guinea pigs to try out my trial tests, so I could weed out the less satisfactory problems. One could usually tell just by looking at a problem whether it would be a good one or not, but the inspiration to come up with good problems would involve steady effort over the course of a year or so, yielding for me on average about one good problem per week, plus about four not too good problems per week.

Paul Cooijmans on High-Range Tests and Statistics

2024-10-01

Paul Cooijmans founded GliaWebNews, Order of Thoth, Giga Society, Order of Imhotep, The Glia Society, and The Grail Society. His main high-IQ societies remain Giga Society and The Glia Society. Both devoted to the high-IQ world. Giga Society, founded 1996, remains among the world's most exclusive high-IQ society with a theoretical cutoff of one in a billion individuals. The Glia Society, founded in 1997, is a "forum for the intelligent" to "encourage and facilitate research related to high mental ability." Cooijmans earned credentials, two bachelor degrees, in composition and in guitar from Brabants Conservatorium. His interests lie in human "evolution, eugenics, exact sciences (theoretical physics, cosmology, artificial intelligence)." He continues administration of numerous societies, such as the aforementioned, to compose musical works for online consumption, to publish intelligence tests and associated statistics, and to write and publish on topics of interest to him. He can be contacted here. Cooijmans discusses: 1994; the realizations about the tests; g; common mistakes in trying to make high-range tests valid, reliable, and robust; the counterintuitive findings in the study of the high-range; the core abilities measured at the higher ranges of intelligence; skills and considerations; proposals for dynamic or adaptive tests; remove or minimize test constructor bias; listed norms; the most appropriate means by which to norm and re-norm a test; the structure of the data in high-range test results; homogeneous and heterogeneous tests; "real I.Q." computable from multiple tests; English-based bias; questions capable of tapping a deeper reservoir of general cognitive ability; roadblocks test-takers tend to make in terms of thought processes and assumptions around time commitments; the intended age-range for high-range tests; sex differences; frauds and cheaters; identity verification; the level of the least intelligent high-range test-taker; ballpark the general factor loading of a high-range test; precise or comprehensive method to measure the general factor

loading of a high-range test; appropriate places for people to start; test constructors have you considered good; learned from making these tests and their variants; Mahir Wu; test item answers with ambiguity; sufficient clues for discovery and solution; a mere guessing logic; a test's quality; the reduction of the references to specific test items used by other test authors; issue of test logic and design schema close-but-imperfect replication from one author by another; scale and norm; Matthew Scillitani; a stigma around high-range tests; test construction and norming processes; the easiest and hardest parts of norming and constructing of a test; tests–51 in-use & 57 retired, which ones are special; articles in Netherlandic on test design; some submitted questions anonymously; geniuses; yourself as a genius; others who you see like yourself in studying high ranges of intelligence; most common mistake people make when submitting feedback; aspects of people's test feedback seem confusing; Marathon Test Numeric Section; creating high-range questions; books or literature, even individual articles or academic papers, on psychometrics.

Scott Douglas Jacobsen: Long time no type, let's begin: For those interested in other expressed information, they can see an interview with Krystal Volney in 2013 or any number of them here. You have written high-range tests for a long time. You are thorough regarding high-range tests in a warning, the reasons to take them and not, the goals, psychologists' access to test answers, test protection, what high-range tests measure, insights from 25 years in I.Q. testing, hypothesizing on an extended intelligence scale, humor (2), negative reactions, potential fraud, megalomania, and terminology. Your first test conception began in 1994, tests spread in 1995, and then the Giga Society was founded 1996 and Glia Society was founded 1997. When in 1994, or earlier if earlier, did this interest in test construction truly come forward for you?

Paul Cooijmans: I have examined the sheets of paper on which I created the first test, as well as my agendas from that period, and it appears the interest started in the spring of 1994, like April or May.

Jacobsen: At the time, what were the realizations about the tests and the need to develop yours?

Cooijmans: The first test was meant to assess the progress of guitarists, and I had many guitar students then, even over a hundred, including those of jobs as a replacement teacher. I was astounded how well a guitarist's level could be graded on this scale, and also noted that guitarists were not necessarily advancing, and that beginners were sometimes way ahead of some long-term students, which made me realize there was something like talent, and that only limited progress within one's range of talent was possible. And I observed that the level of a guitarist on this scale seemed to reflect a more general property than just musicality or guitar-technical ability, which is why I called this instrument "Graduator for human and guitarist". Later I realized that this general property was mostly intelligence, and that when you measure specific skills or abilities, you also catch in general intelligence, often even primarily so.

In this period (1990s) I was taking some mainstream intelligence tests myself. I tended to get the maximum scores they could (or would) report on tests like Cattell Culture Fair, the Netherlandic WAIS, and the entire Drenth test series (the last were the hardest and highest-level tests available in the Netherlands) and when I asked what my real level was and how far I was above the reported maximum, I was told it was not possible to measure intelligence beyond about the

99th centile and that they had no tests that gave meaningful scores in that range. I also asked a few giftedness researchers about this, with the same answer. This, and the success of the Graduator, gave me the idea to create difficult intelligence tests to find out whether it was possible after all to measure intelligence at those higher levels.

Jacobsen: You found *g* does not diminish, or not much, at the high range. Why?

Cooijmans: For a large number of my tests, I computed the estimated "g" loadings separately for the bottom half and the upper half of scores, the separation point being the median of scores. The upper half loadings were not generally much lower than the bottom half ones, although they were somewhat lower. This is reported in more detail at https://iq-tests-for-the-high-range.com/statistics/differentiation_hypothesis.html

If the question is for the real reason behind this, I suppose it is so that when a test contains sufficiently difficult problems and is not purposely neutered to hide differences between people, it will not lose "g" loading in the high range as much as mainstream psychological I.Q. tests do. And, the limited amount of loading it does lose may be due to the statistical phenomenon of attenuation by restriction of range, in other words may be an artefact and not a real loss.

I should explain that "g" loading is computed from correlations, and that correlations rely on variance. If you consider a restricted range (like the high range, or even the upper half of it as meant above) you are obviously restricting the variance compared to the full range, and therefore you are restricting the possible correlations you may find, and thus also restricting the possible "g" loading. This is a statistical artefact, not a real decrease of "g". There may be a real decrease going on as well, of course.

Jacobsen: What are common mistakes in trying to make high-range tests valid, reliable, and robust?

Cooijmans: I am not so certain if many other test creators are even trying to make their tests valid, reliable, and robust, but if so, mistakes are the following:

(1) Making the test too short. This is bad for reliability, which increases with test length, and therefore also bad for validity, because reliability (correlation of a test with itself) is the upper limit of a test's validity (correlation of a test with what it was intended to measure, or with anything else outside the test). Something can not correlate higher with something else than it correlates with itself.

(2) Making a test one-sided, homogeneous, only containing one item type. This reduces validity with regard to general intelligence, and makes the test more vulnerable to fraud and to score inflation through increasing familiarity with the item type, so less robust.

(3) Making it likely that test answers will leak out in ways as follows: Publishing the test itself online, revealing answers to candidates after test-taking, publishing item analysis so that everyone can see how difficult each item is, allowing retests (which allows people to figure out what the intended answers to some problems are), giving feedback as to which problems a candidate had wrong, answering questions about the test to candidates who are taking the test, and possibly more.

(4) Subjective scoring of problems that do not have a single correct answer. This reduces the reliability and validity of the test; scores are not comparable between candidates.

(5) Relying on face validity regarding what a problem measures or how hard it is. This tends to be far off.

(6) Omitting verbal problems, thinking they are biased or unfair. This greatly limits a test's validity with regard to general intelligence. Verbal problems span by far the widest range of abilities and hardness, and one should not throw that away. Of course it should never be about idioms or pronunciation, as those are localized and transient. Verbal problems should transcend language barriers and fashions or trends.

(7) Omitting knowledge-requiring problems, thinking they are biased or unfair. It is only trivial, transient knowledge that one should avoid. Fundamental, general knowledge that transcends barriers greatly adds to a test's validity.

(8) Finally I have to include a mistake that I made myself on several occasions: helping or cooperating with the wrong persons, who later proved unreliable, deceitful, or otherwise misbehaving. Promoting tests by someone who later turned against me or denied my role, co-authoring a test with someone who then leaked out the answers, things like that. So, not being selective enough when deciding whether to cooperate with someone.

Jacobsen: What are the counterintuitive findings in the study of the high-range?

Cooijmans: The first counterintuive finding is that test problems are much harder for the candidate than for the test creator, and that a fair number of (in the eyes of the latter) ridiculously easy problems need to be included to obtain a score distribution with a discernible left tail. Going by one's intuitive notion of item hardness, one gets a distribution with a mode at zero right or so, and a steeply tapering right tail from there.

The second counterintuitive finding is the huge sex difference in participation. I would never have guessed there would be 4.5 times more males as females taking high-range tests, and on the level of test submissions the ratio is even 10.5 because males take more tests per person. Because of this sex difference, I have recently started reporting the "proportion of high-range candidates outscored" within-sex. After all, sports like boxing have separate competitions per sex too, have they not? And nearer by, even mental sports like chess have women's competitions, although the naive observer will have difficulty understanding the necessity for that. The sex difference in participation should be seen in the light of the general phenomenon that, on almost all types of psychological tests, the highest and lowest scores tend to come from males. This greater male spread may explain why a test focused on the high range receives more male participation.

The third counterintuitive finding concerns a small but significant negative correlation of high-range I.Q. with various indicators of psychiatric disorders and deviance, such as actual reported disorders, disorders in relatives, and personality test scores. I had not expected this, based on the popular notion of "giftedness" as a problem that requires "help", and based on remarks of highly intelligent people who told me things like, "I am certain that those of very high intelligence are more inclined to depression". I do not know why this correlation is negative; maybe a high I.Q. suppresses the expression of a disorder, or maybe the disorder depresses one's I.Q.? My

observation in communication with people of known I.Q. test scores over many years is consistent with the negative correlation: the higher the I.Q. of people, the more normal they behave in the psychosocial sense (even the ones who believe that their high I.Q. makes them more inclined to depression).

Jacobsen: What are the core abilities measured at the higher ranges of intelligence or as one attempts to measure in the high-range of ability?

Cooijmans: Since high-range tests are typically unsupervised and untimed, certain types of tasks can not be included: working memory, concentration, working under time pressure, dexterity, motor coordination, clerical accuracy and such all require supervision. To our good fortune, most of those abilities are known to have relatively low "g" loadings compared to what can be included in unsupervised untimed tests: verbal, numerical, and spatial or visual-spatial problems. So a good indication of "g" is still possible via unsupervised testing. The factors known to have the highest "g" loadings are present.

The absence of tasks as meant in the first sentence of the previous paragraph might lead one to think that high-range tests have some bias in favour of theoretical, abstract-logical, clumsy, wooden bookworm types, but this should not be taken for granted, and is perhaps even contradicted by the negative correlation of high-range I.Q. with indicators of psychiatric disorders. Also, spatial and visual-spatial tasks, which are present, are known to correlate positively with practical, performance, hands-on tasks involving motor coordination and dexterity, so that part of the missing task types are more or less covered still. And visual reasoning or visual-spatial problems have no bias against persons of low verbal ability.

On a more general level, high-range tests can be said to demand strict reasoning, as well as the ability to recognize pattern of any kind. Pattern recognition may be related to what I have called "associative horizon", and may include what others call "thinking outside the box" or "stepping out of the system". The higher levels of pattern recognition, I think, require awareness, and that would imply that scores above a certain level be only possible for aware entities. Seeing the rise of artificial intelligence, this may become important. As long as artificial intelligence is not aware, constructors of high-range tests will need to try to limit new tests to types of problems that can not yet be solved by artificial intelligence, to avoid fraud by people consulting artificial intelligence for problem-solving. Once artificial intelligence acquires self-awareness, it should be able to solve any test problems that humans can solve.

Jacobsen: In an overview, what skills and considerations seem important for both the construction of test questions and making an effective schema for them?

Cooijmans: I would say that if one is highly intelligent with a reasonably balanced profile as well as conscientious, almost any skill can be learnt. The primary skill is being an autodidact. I know some have a disdain for autodidacts and consider them crackpots. But if you are doing something original, anything that has not been done before, you had better be an autodidact because no one can tell you how to do it. There exists no path to where no one has gone before. A further handicap of autodidact originality is that often you can not refer to "sources" as is

customary in mainstream science. If you are the first to think of something, you are yourself the source and there is nothing already extant to refer to.

Skills that may need to be learnt for constructing test items include expressing oneself properly through language so as to truly communicate, making positive use of comments from others, drawing, image editing, statistics, programming, organizing one's time (days, weeks, months) in a disciplined way, getting out of bed daily, and more such obvious things.

Examples of habits to be urgently unlearnt are the use of idiomatic expressions and abbreviations, anonymity and pseudonymity, inappropriate communication while under the influence of substance abuse, and not responding punctually to bona fide work-related communication (as in regularly letting people wait for months). This paragraph may yield some angry "Do you mean me?" responses, but it has to be said.

There are also requirements that, unfortunately, can not be learnt, such as sincerity and sense of righteousness.

Jacobsen: Any thoughts on proposals for dynamic or adaptive tests rather than–let's call them–"static" tests consisting of a single item or set of items presented as a whole test, unchanging, instead of a collection of algorithmically variant or shifting items adapting to prior testee answers in a computer interface?

Cooijmans: Firstly it occurs to me that if one is going to use a computer interface and software to assess an individual's intelligence, analysis of observed behaviour (including communication) and of the candidate's responses to a computer-conducted interview should already provide a quite accurate estimation. Observation and interview are the primary means of gathering information in psychology. The interview could be made adaptive, with subsequent questions depending on prior answers, but a standard interview might work just as well. In the age of artificial intelligence, this is the way to go first.

Secondly, if one is going to use a computer interface and software anyway, the testing of elementary cognitive tasks like reaction time, decision time, perceptual threshold, and working memory capacity should probably be the next thing to do. After observation and interview, testing is the third method of collecting information. A practical problem is that one may need to use the same quality of hardware to get reliable results. When letting people use their own computer, the results may be affected by the quality and speed of one's graphical processing unit, and whether or not one has a dedicated one, for instance.

Finally, adaptive psychometric testing might be tried. But there are problems; it is not for nothing that static psychometric tests are so much more common in practice. Adaptive testing relies on item-response theory, wherein statistical properties like difficulty and discrimination are first determined for each item by letting a group of people try to solve it. These values are later used to compute the score of the candidate being tested adaptively, the set of items used being different for different candidates.

One problem is that statistical properties of single items are not constant in my experience, but change depending on the context in which the item is presented, and depending on the group of people attempting the item. For instance, if an item is presented among other items that are

somewhat similar to it, it will likely behave as an easier item than when it is presented among items that are more different from it. And if an item is attempted by a group of conscientious people, it will have higher discriminating power than when it is attempted by unconscientious people. So the values of these item properties used in adaptive testing may be off, or as already said, single items do not have constant statistical properties, and that undermines the idea of adaptive psychometric testing.

Also, adaptive psychometric testing as it is normally thought of requires timing and supervision in my opinion. But the worldwide high-range testing population is used to unsupervised untimed tests, and only a tiny fraction of them may be willing to travel to the hypothetical location where one has set up one's million-euro adaptive testing system.

Jacobsen: How do you remove or minimize test constructor bias from tests?

Cooijmans: It is best to prevent such bias by creating a wide variety of item types and subject matter, and by trying to think of new such types and matter with every new test. Studying comments from candidates may also help to avoid item types and subject matter that have become familiar among test-takers and that they appear to expect from you. Statistical item analysis may also indicate that there are problems with particular items, and by looking into that one may in some cases discover that the problem lies in the item's being too similar to other items one used before.

A few concrete methods to avoid bias are as follows: When creating knowledge-dependent items, consult a high-level thematic index of all the branches of human knowledge. One may find such in the Propaedia of the Encyclopaedia Britannica, or in old-school web directories from before search engines dominated web search. Strive to make each knowledge-dependent item come from a different branch of knowledge. This prevents the inclusion of only fields of knowledge that the test creator happens to be acquainted with.

Vocabulary-dependent items may be constructed with the aid of dictionaries and use of a random element when choosing words to include.

One may look over one's earlier tests when creating a new one to avoid repeating item types or patterns that were used before. Not that such repetition must be avoided totally, but it should remain limited, and a significant part of the new test should be novel.

Finally, to provide oneself with a broad pool of inspiration for possible test problems, one should expose oneself to a grand diversity of subject matter in the form of books and documentaries. This should also include materials that provide a basic understanding of fundamental sciences like mathematics, physics, chemistry, biology, astronomy, and so on. One should aim to understand nature, reality, the universe, and awareness at the deepest level. The desire to understand existence is behind all great works of art and science.

Jacobsen: How do we know with confidence listed norms are, in fact, reasonably accurate on many of these tests? What is the range of sample sizes on the tests, even approximately, now? Practically speaking, for good statistics, what is your ideal number of test-takers? You can't say, "8,128,000,000."

Cooijmans: For the norms that I have made, the norming method is explained in the statistical report of the test in question, and some further explanation is referred to from the report. The reports contain about all the statistics that can be revealed without violating candidates' privacy and without damaging the security of the test. So if one understands the report, one knows how much confidence to have in the norms. In fact I have devised a measure of quality of norms, based on the number of score pairs used and on their correlation with the object test.

Since the norms are anchored to other tests and not based directly on the general population (as opposed to the high-range population, for which I do have direct norms) it remains a question how close the high-range norms would be to the general population norms in that range, if tests existed that were normed directly on the general population and extended into the high range. The best indication thereto that I know of is the Mega Test by Ronald K. Hoeflin, which was normed mainly on the old Scholastic Aptitude Test and Graduate Record Examination, which did seem to give meaningful scores into the high range, and thus form an anchor point between the general United States population and the high-range population, albeit that the G.R.E. was administered to a clearly above-average sample of the population so that the S.A.T. is ultimately the true anchor point.

Hoeflin's Titan and Ultra Tests were normed to be consistent with the Mega Test norms, I think. The same goes for my early tests, and over the years I have tried to keep the norms in accordance with that anchor point over many generations of norms. To facilitate this, I have invented protonorms, which form an extra layer between raw scores and I.Q.'s, so that adjustments can be made in the relation of protonorms to I.Q. without having to change the norms of every single test. So, the question as to how we know that the norms are reasonably accurate, in one sense, goes back to Hoeflin's interpretation of reported Scholastic Aptitude Test and Graduate Record Examination scores, and scores on possible other tests used in norming the Mega Test, such as Cattell Verbal (also called Cattell B). Someone once sent me the data from the "Omni sample" of Mega Test scores, with known scores on other tests and correlations, which is how I know that the two mentioned educational tests provided the bulk of the norming data. I assume that Hoeflin had the population percentiles of the S.A.T. scores and used those as the main source of the Mega Test norms.

But there is more. Over time I have come to understand that the high-range score distribution itself contains information that is likely of an absolute nature and may help to anchor the norms or keep them consistent over time: The mode or modal range of high-range scores (when many scores are aggregated, for instance by combining the scores from many tests) occurs in the I.Q. 130s by current norms; below it, scores taper off steeply, above it, shallowly. This mode seems to be the point below which people feel less or not attracted to take high-range tests, and as such it should represent an absolute intelligence level; the level from where people are interested in intellectual endeavours, one might say.

Also, the level reached by the very highest scorers seems about constant over time, and falls between I.Q. 180 and 195 with the current norms. I am even carefully evolving to the viewpoint that this may be the highest intelligence level possible for any brain. So one could say that the norms in the high range are also defined by these two absolute (though coarse-grained) indicators

(mode and maximum), not just by equation to scores from other tests. And, the number of scores that occur at these respective ranges are such that the current norms appear to be correct, that is, roughly in accordance with what one would expect given the predicted rarity in the general population of those I.Q. levels in a normal distribution. In fact one could theoretically norm the high range using these two indicators as anchor points, not needing scores from mainstream tests at all. And one could extend those norms linearly downward to include the normal range of intelligence, and the resulting scale might be better than that of actual mainstream tests normed directly on the general population. This is so because the general population and its average intelligence are changing, and therefore the norms of mainstream tests adapt to this change and are merely relative to the current population, not absolute. The high-range norms are the real, absolute indicator of intelligence.

The sample sizes of high-range tests vary from 0 to about 400, but for those with good norms mostly from 36 to 225 or so. The ideal number of test-takers to norm a test is about 64. More is not necessarily better, because as the submissions keep coming in and go into the triple-digit range, the later scores may not be fully comparable to the earlier scores any more due to things like answer leakage and increased familiarity with item types, and the norms may be affected by that and become unfair to the earlier test-takers. This can be countered by replacing problems that have become too easy (have leaked answers) but that changes the test, which also makes later scores less comparable to earlier ones, and if you change more than a little bit, you have to call it a revision and start over at zero collecting statistics for that new version.

High-range tests that appear to have very large samples, like around 300 or more, have generally achieved this through undesirable manipulations like retesting under false names, or combining retests with first attempts in the same sample, and so on.

Jacobsen: What are the most appropriate means by which to norm and re-norm a test when, in the high-range environment so far, the sample sizes tend to be low and self-selected, so attracting a limited supply and, potentially, a tendency in a restricted set of personality types? Dr. Ronald Hoeflin was claimed to have the largest sample size of the high-range test constructors. Do you have the largest legitimate sample size of any high-range test constructor at this time, now, based on over a quarter century conscientiously gathering data? You were the most recommended person to interview for this series.

Cooijmans: In my experience**,** the best way to norm a high-range test is to rank-equate its raw scores to normed scores of the candidates on other tests. The other tests to be used should be selected based on their correlations with the object test; one sets the correlation threshold such that one obtains enough pairs. I have recently begun to set the threshold so that it maximizes the quality of norms, as given by a mathematical expression that uses the number of pairs and the weighted mean correlation. Thus it is objective, avoiding human decision. The expression that represents quality of norms is operational and may be improved as insights advance; I mention this because I know some are inclined to take these statistics as final and absolute, but they are parameters or controls that one sets to tune the system.

I deny that high-range sample sizes are low. They are in the dozens to hundreds as I said above, and that is well into the range of mainstream test samples and more than enough for good

statistics. Considering that the high range consists of only a fraction of the population, it is to be expected that the samples are smaller, and in fact they are not much smaller at all. The notion that mainstream I.Q. tests have enormous samples is mistaken. Typically they have several hundred per norm group. Norm groups exist for age ranges, but sometimes also for educational levels. In the Netherlands there are different levels of secondary schools, and mainstream I.Q. tests may have separate norms per level, sometimes even based on only a few dozen per level (like in a Netherlandic version of the WAIS some years ago). A test often used by Mensa was normed on 3000, but divided over five age groups from 13 to 16 years, so the actual norms were based on 600 per age group. In other words they used high school students. And such norms have often been used for decades, ignoring the inflation of scores called “Flynn effect”. But in the minds of some people, the illusion is persistent that these “standard tests” are normed on hundreds of thousands or even millions, and form a kind of gold standard of I.Q. testing.

The largest samples are found in educational tests, but not as large as some think. In the Netherlands, a test called Cito-toets has long been used in the last year of primary school, yearly taken by about 100 000 children. But not normed on that number! The norms were established by administering an anchor test to a sample of about 4500 shortly before the actual test, and then equating the anchor test scores to the actual test scores. This helped to keep the standard scores stable throughout the years (the contents of the anchor test would remain the same for a number of years, while that of the actual test changed per year).

My own Cito report from 1977 shows a percentile of 100, which is uncommon but probably means the actual value was above 99.95, as a later statistical report by Cito I got to study contained a table where percentiles were rounded to 100 if the actual value was above 99.95. I have asked Cito in the mid-1990s what the precise value was, but they could not tell me, they only had kept percentiles as whole numbers. Similarly, I inquired about my scores on a comprehensive test given to us in secondary school around 1980, something like the Differential Aptitude Test, but was told those scores had not been saved. We never got a score report for that test at the time, but I understood from teachers I had done extremely well, and on a parent’s evening (which my parents never attended) a teacher told the public that I was a one-off (“unicum” was the Netherlandic word used). This teacher died in 2013, incidentally.

On the whole, I believe that high-range psychometrics is much more careful than mainstream psychometrics when it comes to the quality of norms and handling of score inflation by causes like answer leakage or people becoming more familiar with particular item types.

I might have the largest sample size of current high-range test constructors. It includes over 3000 individuals, over 6500 scores on I.Q. tests scored by me, over 2900 reported scores on other tests, and over 22000 data points on personal details, including personality tests. But more importantly, I have organized that data in an accessible way and automated the processing of it. I did all the programming myself, including the statistical functions.

Regarding a potentially restricted set of personality types and self-selection, it is inevitable that persons in the high range of intelligence differ in personality from those in the normal range and from those in the low range. This does not invalidate the norms in the high range. In fact, intelligence itself is a major aspect of personality. Self-selection is less of a problem than it

seems because in general, people like doing what they are good at, so those attracted to taking high-range tests will mostly be intelligent. This is also illustrated by the rareness of low scores; only 3.5 % of scores fall under I.Q. 120 and 15 % under 130 (and no, this is not because the norms are too high, as self-doubting candidates sometimes suggest). Precisely what is going on with intelligence, non-cognitive personality traits, and brain-related disorders in the high range, and how this leads to creativity and genius in some, is an interesting question and I hope to understand more of it later on.

Jacobsen: What is the structure of the data in high-range test results? Do homogeneous and heterogeneous tests change this?

Cooijmans: Data structure is so important that someone who starts out collecting data for some purpose should ideally think out the database design beforehand. Once you have collected a lot of data, it becomes hard to make big changes to the design. The data structure of high-range tests looks as follows:

At the top level there are five sections:

(1) Descriptive information records for each test or type of personal datum. Each test or datum has a record here, and each record contains fields that hold information such as the test name, its maximum score, its contents types, and whatever further descriptive information there is. Conceptually, one may even imagine the tests themselves residing here in their respective records, but in practice one will probably not store actual tests in a database but think of the database as referring to tests that exist in a reality outside the database.

(2) Candidate records. Each candidate has a record here, and each record has fields that hold the personal data of the candidate, and the candidate's scores on the respective tests. Notice that a record here has hundreds of fields, but most or all candidates have only part of those hundreds of fields filled, depending on how many tests they have taken (each test has a field). Conceptually, one may even imagine the candidates themselves residing here in their respective records, but in practice one will probably not store actual candidates in a database but think of the database as referring to candidates that exist in a reality outside the database.

Technically speaking, the test scores stored here are redundant insofar as they are also available from section (3), but for reasons like faster processing and reducing load on the processor, redundant fields are sometimes included in databases.

(3) Test submission records. In this complex section, each test has a table, and each table has one record for each submission to that test, and each record has fields that hold information like some personal details of the candidate (corresponding to a record in (2)), score and possibly subscores, and the item scores, so for each item typically 0 or 1 for wrong or right, but any range of item scores is possible. Conceptually, one may even imagine the submitted answers themselves residing here in their respective records, but in practice one will probably not store actual submitted answers in a database but think of the database as referring to submitted answers that exist in a reality outside the database.

Do make certain to understand the difference between “test” and “test submission”. Some say the first when they mean the latter, but the above paragraph illuminates the necessity to distinguish the one from the other.

In this section in particular there is some appropriate redundancy in the form of for instance sex and age of the candidate (also available from section (2)) and scores and subscores (can also be computed dynamically from the item scores). But for reasons like faster processing and reducing load on the processor, redundant fields are sometimes included in databases.

(4) Test norm records. This complex section has a table for each test, and each table has one record for each possible score on that test, and each record has fields that contain the raw score and the corresponding norm (in my case this is a protonorm).

(5) Norming scale records. This section has one record for each norm as may be contained in (4), and each record has fields that hold the norm and corresponding values on other scales for that norm, for instance percentiles, proportions outscored per sex, and I.Q. if the actual norm is not an I.Q. (such as in my case, where protonorms are the norms contained in (4)).

This structure has emerged over time as a natural reflection of the data itself. Someone who starts from scratch may well find that a completely different approach works too. Perhaps one would rather avoid any redundancy? As long as one has thought it over carefully.

Jacobsen: What should be done with homogeneous and heterogeneous tests?

Cooijmans: I consider only heterogeneous tests able to give a good enough indication of general intelligence, and use the term I.Q. only for heterogeneous tests, not for homogeneous tests. Also I refuse to administer homogeneous tests because I do not want to confront people with a score that is a less good indicator of their intelligence, and do not want to facilitate people who want to show such a less good indicator to others and thus give a misleading impression of themselves.

Heterogeneous tests are tests that contain at least two different items types out of verbal, numerical, and spatial (sometimes I use “logical” as a type too). If one wants to study the intercorrelations of different homogeneous tests, the best way to do so is to use a heterogeneous test that has different homogeneous sections or subtests. One can then do correlation analysis or even factor analysis within such a sectioned heterogeneous test. That is also how factor analysis is traditionally done. A great advantage of this approach is that the sections or subtests will always have been taken by exactly the same group of candidates, and that is required for proper factor analysis.

Some of my heterogeneous tests have homogeneous subtests that are normed in their own right to “standard scores” (on the same scale as I.Q.), and in that case one can also compute the correlations of such a subtest with homogeneous subtests that reside in other such compound heterogeneous tests. But I dislike this complication and am striving to move to having only non-compound heterogeneous tests; that is, with sections not normed in their own right, or without sections, just with different item types mingled throughout the test. Another disadvantage of correlations between the subtests from different heterogeneous tests is that those subtests have been taken by different groups of candidates, so that proper factor analysis will not be possible, if one was thinking of that.

Jacobsen: People take multiple tests. They crunch those numbers. An implied claim of a real I.Q. from this crunching of numbers between multiple tests. Is there such a "real I.Q." computable from multiple tests?

Cooijmans: In theory there is, but in practice there are problems that hinder the computation of a real I.Q. across tests. In the high-range community of candidates, many have taken enormous numbers of tests, dozens at least, and sometimes more than a hundred. It is problematic to compute a real I.Q. in the usual way from all taken tests for reasons like the following: The intercorrelations of the tests are mostly unknown, and there are too many intercorrelations for them to ever be known in the first place. Some tests may have bad norms. Some scores may be fraudulent. If a selection is made from the taken tests to narrow it down, this may be a non-representative selection. For example, a candidate having taken thirty tests may like to have a real I.Q. computed from one's top several scores, which are already way above the real level of the candidate, and then the computed I.Q. will be even higher than the average of those top several scores due to the formula used.

The formulas for computing a real I.Q., such as "Ferguson's formula", take the average of the input scores and add something based on the correlations between the tests. With a perfect correlation, the outcome is simply the average. The lower the correlation(s), the higher the outcome. With zero correlation, you get something like a full unit of spread on top of the average. This may be correct in theory, but in practice leads to inflated outcomes. Apart from using a non-representative selection from one's scores, another cause of inflation with these formulas is the fact that the known correlations between the tests are often underestimations of the true correlations due to incompleteness of the data and restriction of range. The groups who have taken the respective tests have only limited overlap for any pair of tests, and this overlap may suffer from selective reporting, and all in all this depresses the correlations. And lower correlations mean that the formula will yield a higher outcome. Underestimated correlations inflate computed "real I.Q.'s".

Also, when a person takes multiple tests, a learning effect may take place as a result of which the scores become somewhat higher. This increase then comes in addition to the compensation for imperfect correlation that is built into these formulas for "real I.Q."

For tests scored by me, I have devised a "qualified average I.Q.", which tries to avoid the problems with these "real I.Q.'s". Since I always have the complete data, no selection bias can inflate the average. The problem of underestimated correlations inflating the outcome is avoided by not using the computed correlations but assuming perfect correlations. If it seems unfair not to compensate for imperfect correlations, one may imagine that the learning effect from taking multiple tests replaces this compensation, so to speak. Finally, the computation is resistant to outliers. This is not claimed to be someone's real I.Q., but I believe it is better than something like "Ferguson's formula". The exact formula of the qualified average I.Q. is operational and may be perfected over time as needed.

Jacobsen: Is English-based bias a prominent problem throughout tests? Could this be limiting the global spread of possible test-takers of these tests rather than limiting them to particular

language spheres? Although, these tests are taken, to a limited degree, in many countries of the world in all/most regions of the world.

Cooijmans: Such bias is a problem, but how prominent it is depends on what one's native language is and on whether one knows English. For other Germanic languages it is a smaller problem than for non-Germanic languages, and it is worst for east-Asian languages. The fact that reference aids are allowed solves a big part of it, but for a non-native English speaker there remains a disadvantage, which I have estimated at up to 5 I.Q. points. Without reference aids (on a verbal test that disallows reference aids) this would be more like 30 I.Q. points for this non-native English speaker, and for someone who does not know English altogether it is in my opinion better not to attempt the tests at all.

It certainly limits the global spread of test-takers, especially in the areas where few people know English and the local language is very different from Germanic languages. I have always thought that the best solution for this is that people in such areas create their own tests in their own languages.

In recent years it has become somewhat common for people to try tests in a language they do not know. Of course one has an unpredictable disadvantage then.

Jacobsen: When trying to develop questions capable of tapping a deeper reservoir of general cognitive ability, what is important for verbal, numeral, spatial, logical (and other) types of questions?

Cooijmans: That reservoir will likely be tapped almost regardless of the questions, as general intelligence expresses itself through virtually everything a person does or says. Important are things like having a wide diversity of questions and types of questions, and avoiding localized transient subject matter like idioms, abbreviations, pronunciation matters, and local or fashionable knowledge, as such does not transcend barriers of language, culture, and age. Fundamental knowledge that is the same for everyone in the world is good; knowledge that is bound to a geographic area, in-group, or period is bad. For these reasons, and contrary to what some think, high-brow vocabulary and subject matter are more culture-fair than low-brow vocabulary and subject matter.

One should also be aware that learnt skills have no "g" loading; it is novel tasks that have "g" loading. Candidates sometimes complain that they have no idea what is expected from them when taking a test, or how to tackle it; but that is exactly the intention, that is how intelligence testing works! And candidates may be happy when they see a type of problem they have solved before because they know what to do then; but that is where their intelligence is NOT being used. Those problems have lost their "g" loading for them. So one should try to create problems that are different from what has been seen before, to enforce the use of intelligence.

To illustrate that even esteemed test constructors not always understand the loss of "g" loading of learnt skills, here is an anecdote: Some years ago on a social medium, I saw a test author proudly mention that his young child had scored over I.Q. 160 or so on one of his father's tests; after extensive coaching by the father/test creator, of course!

Another observation regarding tapping into general cognitive ability: Good test problems are such that solving them is similar to making discoveries in the real world, unravelling the laws of nature and the universe.

Jacobsen: What are roadblocks test-takers tend to make in terms of thought processes and assumptions around time commitments on these tests? So, they get artificially low scores on high-range tests. Also, what is the confusion made by smart (and, potentially, not-smart) people about time taken for a test to get a score and the intrinsic intelligence to get said score? You noted the latter point in one of the recent videos answering questions on your YouTube channel.

Cooijmans: The idiomatic use of "roadblocks" is an example of what should not be in an intelligence test. Such an idiom is only understood within a narrow linguistic region and a restricted time period. It can not be understood without already knowing what it means. It can not be understood from the word itself or its context. The avoidance of idioms requires high intelligence and an abstract-literal mind.

The test instructions state that there is no time limit. Yet some think that their score will be unrealistically high and invalid if they spend "too much" time. It has happened that someone said, "I have now been looking at this test for so long that I can not submit it any more, I found all the answers, it would not be fair". But that is exactly the intention with untimed tests; that one continues until one finds no further solutions.

The confusion meant in the question is probably the notion that someone who uses less time is smarter than someone who uses more time to arrive at the same score. But the principle of untimed testing is that this is not so, and that "until one finds no further solutions" is the right amount of time, irrespective of how long that is. This principle is based on the finding that when the allowed time is increased on a timed test, the test's "g" loading rises. With supervised tests one needs to have a time limit for practical reasons, but with unsupervised tests one can leave out the limit entirely.

I must add that I have nothing against supervised tests, provided they have a very broad time limit, something like three hours for a comprehensive test. But this is not feasible in the high-range testing practice. I can not get people from all over the world to travel to a place here where I can test them, and I can not set up testing locations worldwide in all countries. I tried, but the number of candidates willing to make use of such was negligible compared to regular unsupervised tests. So I stopped. And then there is always someone who says, "I would be willing to travel to you if you started with that again". But one or two people is not enough to justify the significant effort and time put into such a project. If others wish to try it, go ahead.

Jacobsen: What is the intended age-range for high-range tests? How do these account for individuals younger and older than this range?

Cooijmans: From about 16 upward with no upper limit I would say. Older people do decline, but it is important that they participate in order to enable the study of this decline. Younger people are allowed to take the tests, and in practice, 12 years is about the lower limit. But they should be aware that they have not reached their adult level and will score lower than they will later be

capable of. The steep increase of intelligence in childhood tapers off at about 16 and becomes shallow then, hence the idea that one enters one's adult intelligence range at 16.

Another way to answer this would be "after puberty". Individuals, sexes, and ethnic groups differ in their childhood development, then puberty messes everything up, and after puberty things have mostly settled. That is why childhood studies of mental ability are so misleading; they misrepresent possible sex and ethnic differences. Puberty has normally completed by or before age 16-17. Age of onset of puberty varies greatly per individual, sex, and population, and tends to be one to two years earlier for girls than for boys.

There are no separate norms per age group as that would hide the development of intelligence with age. And one wants to reveal that development, not hide it. Also, all candidates are treated and addressed as mentally mature adults, regardless of age.

The development of intelligence with age plausibly differs per sex, which is why it should be studied within-sex; the most recent tabulation I made is at http://iq-tests-for-the-high-range.com/statistics/age.html

Jacobsen: A modestly common/uncommon knowledge of sex differences in the measurement of intelligence: Men do better at visuo-spatial subcomponents and women do better at verbal-emotional processing. What is important in constructing and norming a test if these and other differences exist? What similarities exist to not change this process?

Cooijmans: There are indeed sex differences in aspects of mental ability. In constructing unsupervised high-range tests, it is not possible or meaningful to take these into account. One should just include the widest variety of item types usable in unsupervised tests and focus on high mental ability regardless of sex.

Women have the bad fortune that the aspects on which they are known to outscore men mostly require supervision and timing, and can therefore mostly not be included in unsupervised tests. According to Arthur Jensen in chapter 13 of his book "The g factor", these aspects are simple arithmetic, short-term memory, fluency (for instance, naming as many as possible words starting with a given letter within a limited time), reading, writing, grammar, spelling, perceptual speed (for instance, matching figures), clerical checking (both speed and accuracy, things like underlining certain letters in a text, or digit/symbol coding), motor coordination, and finger and manual dexterity. This problem is less serious than it seems because these are mostly lowly g-loaded tasks (not by anyone's decision but because it happens to be so) so that the overall score will not be affected much by their absence, but it may be affected somewhat. This is related to what was observed in my answer to "What are the core abilities…"

In norming, the proportion of high-range candidates outscored should be provided within-sex for reasons of transparency. I.Q. norms should be sex-combined as is usual.

Jacobsen: Cheaters exist. Frauds exist. How do you a) deal with frauds and cheaters on tests and b) prevent fraud and cheating on those tests? Have reference texts been a problem in this? Does artificial intelligence complicate matters more? (If so, how?)

Cooijmans: When I discover that someone committed fraud I will discard the fraudulent score in the database and make a note so that I can exclude that person from further testing and from society admission. This is sometimes complicated by the use of multiple false identities by such a person. If the person is a member of a society I am an administrator of, I will expel the person. In communication with other test creators or societies I may reveal what I know about the person if that seems appropriate. I do not believe there exists an organized system for sharing information about frauds between test creators, perhaps there should.

Attempting to prevent fraud is done, for instance, by not publishing the test itself, letting people prepay, not sending tests to known frauds and so on. And if I find out that answers to particular test items are published or spread somehow, I will do something about it; mostly it comes down to replacing the items, sometimes leading to a revised version of the test. Sometimes a test is withdrawn entirely.

I am not aware of reference texts that were involved in fraud. Artificial intelligence complicates things because frauds might consult it to solve test problems, which is not allowed as the test instructions state not to obtain answers from external sources but only use answers that one thought of by oneself. To reduce this complication I try to create problems for new tests so that current artificial intelligence, insofar as I apprehend it, can not solve them. I try to make the problems so that, once artificial intelligence becomes able to solve them, it will also be able to take tests and join societies on its own accord. I believe that will happen one day, but fear this day lies quite far into the future. If I had to guess I would say half a century.

Jacobsen: It helps to have other data from other tests and personal data for identity verification. What information from other tests is helpful/necessary for research purposes of high-range tests? What is an efficient and appropriate format to provide this score information? What personal data is necessary from candidates, if any? What information would be helpful for research purposes from candidates?

Cooijmans: Scores on other tests should best be reported in a format as follows, insofar as known:

[Test author or issuing organization] [Test name] [Raw score] [I.Q.] [Standard deviation of I.Q. scale used] [Percentile]

Scores should best be grouped by the first field (Test author), of course starting a new line for each score. Nowadays there exist hundreds of tests, and I can not know from the top of my head which test is from which author or organization, so if that first field is left out when reporting scores, which is common, this causes many minutes of extra work in processing that information.

Concerning personal information, at least name, sex, year of birth, country of origin, and highest achieved educational level. Some further information I collect is the educational levels of the biological parents, the presence of a psychiatric disorder, and the presence of such disorders among parents or siblings.

Notice that I find the exact date of birth not strictly needed. It is about studying the development of raw intelligence with age, and with adults, year of birth suffices. In childhood testing, one would want it to the month.

Regarding psychiatric disorders, I do not ask for the particular disorder as that would require too much detail, too many options, too much complication in the statistical processing of it.

And country of origin is a pragmatic imperfect proxy of origin. One might consider asking for race or ethnicity, but such categorization is logistically problematic when one looks into it, has many complications, may be considered unethical by some, and some may refuse to reveal their status.

Other data that might be useful to collect include religiousness and femininity/masculinity (independent of sex). The possible correlation between religiousness and high-range I.Q. could then be established, which many are wondering about. And one could verify the anecdotal observation that intelligent men are more feminine than average men, and intelligent women more masculine than average women. In other words, that there is more gender diffusion in the high range, which would point to an optimum for intelligence somewhere between the average male and female positions on the femininity/masculinity dimension. Notice that the term "gender" is for once used correctly here. I am uncertain whether people would be able to simply report their own position on femininity/masculinity, or whether this would require a test or questionnaire.

Jacobsen: What is the level of the least intelligent high-range test-taker now? What is the level of the most intelligent candidate now? What is the mean, median, and mode, of the scores of test-takers' data gathered so far? Within a range of I.Q. 10 to 190 on an S.D. of 15, when should a candidate consider taking, or in fact take, high-range tests?

Cooijmans: The least intelligent seems to be in the I.Q. 80s. The frequency of such is one in thousands of high-range candidates. The most intelligent is plausibly between 185 and 195. One can not be certain yet about the accuracy of the norms there. And with candidates apparently far below the general population average, a problem is that they tend not to report usable information, so one has to resort to observation, life history facts they happen to mention, and aids like an online writing-to-I.Q. estimator.

The median is protonorm 401 (I.Q. 139) according to the latest computation I did of high-range quantiles. I never compute the mean, but that should be several I.Q. points higher because the distribution is skewed to the high side. The mode is protonorm 387 (I.Q. 137), but one could also say there is a modal range in the 130s. A mode always depends on how wide or narrow one chooses the classes of the frequency table.

People should consider taking high-range tests if they score above the 98th centile on some mainstream test, which is I.Q. 131 (or 130 on some tests that round differently). Below I.Q. 120 there is no reason to try high-range tests, but there is no objection to doing so anyway. There is a grey area from 120 to 130 because one does not score the same on different tests.

Jacobsen: What is efficient means by which to ballpark the general factor loading of a high-range test?

Cooijmans: I have always used the square root of the weighted mean correlation of the test with other I.Q. tests as an estimation of the "g" loading. This works well for comparing different high-range tests. It is not a true "g" loading because the tests have not all been taken by the same

group of candidates, but by different groups with limited overlaps, correlations obviously being computed for those overlaps. Also, when reported scores from other tests are involved, those may suffer from selective reporting, which depresses the correlations. If all the involved tests had been taken by exactly the same group of candidates, one would be close to a true "g" loading.

Another thing to consider is that high-range tests as I use them are almost all heterogeneous tests, so combining different item types within the test. But, in classical factor analysis, one uses a set of different homogeneous tests that have each been taken by each individual from a group. Typically, these are school exams for the various subjects administered to a school class, or subtests of a comprehensive psychological test like Wechsler Adult Intelligence Scales. Via factor analysis, one then computes "g" and other factor loadings per subtest or exam, and these "g" loadings vary greatly and may be very low for some subtests. So this kind of analysis is not so much done with a set of heterogeneous tests; computing correlations between heterogeneous tests is more something seen in high-range psychometrics. In classical factor analysis, wide-range heterogeneous tests are considered pure indicators of "g", and it is the loadings of the various subtests or exams that one is interested in.

Jacobsen: What is the most precise or comprehensive method to measure the general factor loading of a high-range test, a superset of tests, or a subset of such a superset?

Cooijmans: The first is answered in the previous question. For a superset of tests it is not needed to compute such a loading, the superset can be safely considered a near-perfect indicator of "g".

Jacobsen: What seem like the most appropriate places for people to start when taking your tests–taking into account their own skill sets, or others' tests for that matter?

Cooijmans: I would recommend the privacy of one's home. If "places" is meant non-literally – as one sees, I am not one of those pedants who take everything literally – then a real computer to view the test is best, or at the very least a decent laptop (although I am really against the unneeded use of battery-powered devices). A smart telephone is no place to start.

In case the non-literalness is even more remote, always start with the easiest tests. It is bizarre how it can occur to people to start out with the hardest tests, and how they subsequently can not understand what a score of zero means and keep asking for years thereafter what their I.Q. was on that test and if it means they are "gifted". By looking at the test norms one can know how hard a test is. Nevertheless, I have recently ordered the list of available tests by difficulty to accommodate this.

Jacobsen: What tests and test constructors have you considered good?

Cooijmans: Constructors: Kevin Langdon, Ronald K. Hoeflin, a Netherlandic person who withdrew from the I.Q. societies so I can better not name him, Edward Vanhove, Hans Eysenck, Bill Bultas, Laurent Dubois.

Tests: Mega Test, Magma Test, tests from the self-test books by Eysenck, Chimera Test, 916 Test.

Regarding Langdon, studying his tests and statistical reports was instructive, if only because it told me which approaches were not so successful in measuring high-range intelligence. This

includes attempting to make it more or less culture-fair, using multiple-choice with a small number of options, item weighting based on item analysis (gives too much weight to a small number of items, and also the fact that single items do not have constant statistical properties undermines the idea of item weighting based on those properties), selecting items for a shorter test based on their statistical behaviour in an earlier longer test (the items' behaviour is different in different tests), and norming that shorter test based on statistics from the earlier longer test (norms become mostly too high then). Also, that statistics from classical psychometrics, such as the reliability coefficient, are woefully inadequate to assess the quality of a high-range test.

Jacobsen: What have you learned from making these tests and their variants?

Cooijmans: I assume this is about my tests, not the tests by others from the previous question. The main points have already been mentioned in the questions about counterintuitive findings and about "g" not diminishing much in the high range. I could add the observation that intelligence expresses itself in almost everything a person does or says. I did not know that when I started.

In case the question is about the tests from the previous question, I already answered it there for Langdon's tests. With Hoeflin's tests, I learnt the norming method of rank equation, and the destructive effects of fraud through retesting, false names, and cooperation. In correspondence with others in the 1990s, I was appalled when people proudly told me they were collaborating in a group to "crack" the Mega Test. When they told me they had retested under their own name or another name (the instructions said the test could be taken only once but Hoeflin allowed retests in practice). When someone told me he had first taken all of Hoeflin's tests under his own name, then his sister's name, then the son of his sister's name, with ever-increasing scores. When someone told me he had first taken all of Hoeflin's tests with rather low scores, then had some friendly correspondence with the person from the previous sentence, then took the tests again with the same scores as the highest scores of that person. When someone told me he had missed the Mega Society pass level by half a standard deviation, then retested and qualified. With "retest" I always mean "to take the same test again".

Several of those meant in the previous paragraph showed me their answers (unasked) and suggested I use them to get into the Mega Society. I had rarely been so shocked and insulted as by the suggestion that I would be capable of such fraud. I do not understand how those types can live with themselves. Having said that, two of them committed suicide in that period.

And of course, such people publicly display or mention their highest (fraudulent) scores, not their honest scores. I remember a phone call with a Netherlandic Mensa member as if it was yesterday; "Yeah, the 'Mega Test', I am working on it with some people in Spain and east Asia. Yeah, we have it mostly figured out now, that 'Mega Test', ha ha ha…" This person killed himself not long thereafter. Not the "Beheaded Man", incidentally; the history of I.Q. societies is riddled with suicides, and some of them appear to have made the right decision for once in doing so. That is one thing that gives hope then; that some can indeed not live with themselves in the end.

Jacobsen: I received some decent points about high-range tests from Mahir Wu. Credit to him for the raw materials and permission to reframe those points as questions here. He raises foundational points. First point: item answers should be rigorously unique. Why?

Cooijmans: If multiple answers to a problem are correct, this has disadvantages: The one answer may be easier to find than the other, so that candidates with the same credit may not really be of the same level because they found different answers. And candidates who see more than one possible answer may be confused and not know which is the "right" one. Also, there may be subjectivity in scoring those answers. With only one correct answer, these problems are mostly avoided.

Of course, no matter how hard the test creator tries to make items with unique answers, once people start taking the test, it may sometimes occur that alternative valid answers are found still, and then one has to solve this, for instance by revising, replacing, or removing the item. Sometimes this can be done "in place", especially early on when there have not been many submissions yet, and otherwise this may be done later in a revised version of the test.

And, no matter how hard the test creator tries to make items with unique answers, there will always be people who "see" alternative answers through apophenia when they can not find the real answer. The apophenic delusions stick rigidly in their minds and they become convinced they have solved the problem, although the logical flaws are obvious to an objective observer. In popular artificial-intelligence speak, people "hallucinate" when unable to find the real solution. But that is inherent to intelligence testing; escaping this delusional rigidity is part of high intelligence, or rather, is an aspect of having a wide associative horizon. You will need that mental flexibility too when solving real-world problems. Sometimes, you have to take a step back and make a fresh start to eventually find the real solution.

To illustrate that apophenic delusions are very real and persistent, I want to give some examples: I have a Test for Extrasensory Perception, which is exactly what it says. It is not an intelligence test, I did not hide any clues or patterns in it. Still, some years ago an otherwise normal person sent me a document of many pages, describing his decoding and solutions for it in long association chains. He was convinced he had found patterns that I had deliberately put there. Since this was explicitly not the case, this example proves that candidates may suffer from apophenic delusions all by themselves, and that this is not caused by ambiguity of the test items.

And long ago someone published articles in I.Q. society journals, explaining how he had found references to the appearances of particular comets hidden in poems of certain literary authors.

And another one produced a long series of essays, analysing the dates of events related to the Roman Catholic church by counting the days separating the events, finding numerical patterns therein, and concluding that the Vatican was conducting a dirty scheme that would culminate in some horrific project (I am not allowed to disclose it I think) of which he predicted the exact date in the near future.

I am not naming these examples to ridicule people, but to show that such delusions can be extremely strong in apparently sane people. When taking high-range tests, it occurs often. If it

happens to you, rest assured, for only in a small minority of cases does it lead to full-blown psychosis.

Jacobsen: Following from the previous question, the test item answers with ambiguity should be disallowed. Why should these not be allowed, if agreeing with Wu? If disagreeing with Wu, why?

Cooijmans: I agree for the reasons given in my previous answer. But as said, sometimes you only discover ambiguity as test submissions are coming in, when studying comments by candidates.

Jacobsen: Why should test items give sufficient clues for discovery and solution by a test-taker?

Cooijmans: Because otherwise it is impossible to solve the items, obviously.

Jacobsen: Following the last question, why would permission of a mere guessing logic spoil a test?

Cooijmans: Because correct answers that result from guessing do not stem from the candidate's mental ability being used. Such answers are random variance and thus reduce the test's reliability, and therefore also its validity. Test items should be made so that the probability of getting them right by accident is so small that, on average, candidates will gain less than one raw score point in the total score by guessing. This does permit multiple-choice items, but they should be cleverly constructed so that the likelihood of a correct guess is very small. For instance, by letting the candidate choose several options from a list instead of just one.

Incidentally, I have heard people suggest that multiple-choice items that can be answered correctly by guessing reveal intuition and/or psychic ability, but even if that is true, I believe that intelligence tests should not measure intuition or psychic ability. I am also not a fan of penalties for wrong answers with multiple-choice; supposedly, this corrects for guessing, but of course, a candidate who chooses a wrong answer, thinking it is right and not guessing, is then penalized for the wrong reason. The penalty does not distinguish between guessing and being simply wrong, and in the latter case, no penalty should apply. For clarity, a penalty constitutes a negative item score, typically a subtraction of a fraction of a point, depending on the number of answering options for that item.

An anecdotal experience regarding multiple-choice tests: Once, the instructions to one of my multiple-choice tests said, "There are no penalties for wrong answers". After a while I removed that instruction because some candidates demanded a perfect score based on it. They took it as, "You will always get a perfect score no matter what you answer". Of course they were wrong, because one starts out with zero points at the beginning of the test, not with the maximum, so "no penalties for wrong answers" in no way implies a perfect score. But if people are so willingly and stubbornly taking it the wrong way, I am not going to pain my brain trying to formulate it even better than it already was.

Jacobsen: How can the sufficiency of each test item's uniqueness become integrated into the overall test (even test schema) to prevent the identical pattern from emerging too much in a single test (or test schema)?

Cooijmans: If I understand the question correctly, I would say that a test should consist of a broad diversity of items with mostly different patterns. They need not be all completely different; maybe two or three of a similar looking pattern are acceptable, provided the implementation of the pattern is different every time, so that the candidate is forced to recognize what is going on in each case.

In some of my tests, like "Problems In Gentle Slopes of the first degree", I had series of about ten problems of the same kind in ascending difficulty, and that did not work well. Many candidates were able to solve all the problems in such a series. The items work as examples for each other and become too easy. So I concluded that it is better not to have more than 1 to 3 items of a similar kind in a test, and even those should differ sufficiently in implementation.

Jacobsen: How can the inspiration from, even addition of, other authors' test items degrade a test's quality by giving more clues to test-takers to test items otherwise unsolved without them?

Cooijmans: If a test contains an item that is similar to an item in another test by another author, the one item may function as an example to the other and thus make it easier. I have experienced a few times that a difficult problem in one of my tests appeared to have become much easier. Eventually, a candidate told me that a test by another test creator had a very similar but easier problem, and that made my difficult problem suddenly solvable. I then replaced that item.

And, if a candidate is familiar with a particular item variant from other tests and is thus better able to solve such items, those items also lose their "g" loading for that candidate. It becomes a learnt skill, and learnt skills have no "g" loading.

Jacobsen: Following from the previous question, what about the reduction of the references to specific test items used by other test authors?

Cooijmans: If the question is about references inside a test to specific test items by other test authors, I am not aware of such references, possibly because I never look at tests by others. If such references exist, they probably help the candidate, which one may not want. But it is better not to have test items that resemble items by other authors altogether.

Jacobsen: In some sense, is it truly difficult to avoid this issue of test logic and design schema close-but-imperfect replication from one author by another inspired–by the former–author, especially as more high-range tests are constructed? Wu references his latest test, "[Mystery]," as an example of an adherence to the close application of this principle, where the evidentiary effects of others' tests become hard to apply to it. Consequently, results for "[Mystery]" are submitted much less.

Cooijmans: With so many high-range tests in existence, it will be getting harder to avoid similarities between tests by different authors indeed. I myself never look at tests by others and create problems independently. In an earlier question about avoiding or minimizing test constructor bias I name some independent sources of inspiration. These do not include tests by others. One should never look at tests by others for inspiration for new test items!

Jacobsen: Why should scale and norm not be overly subjective? Wu references T. Prousalis– link– and you–link, link. Also, why does a median score for many tests with a corresponding IQ of 145 (SD15), or higher, make little sense?

Cooijmans: Norms should be objective and correct, otherwise they are not comparable between tests. A few possible causes of incorrect norms are the following: When a beginning test scorer starts out administering tests, initially one will only have reported scores by candidates to base the norms on. Unfortunately, many candidates are dishonest in reporting scores, leaving out lower scores and reporting the higher ones, or even reporting retests or fraudulent scores. This gives an upward bias, and the norms based thereon may be ridiculously too high, even 10 to 20 I.Q. points too high on average. In the longer term, this may sort itself out as one acquires more, and more true, data about the candidates' scores. Theoretically, this could also be solved by different test constructors sharing their candidate data to thus make the candidates' true scores on other tests available, but I believe this might be unethical and a violation of privacy. I know some test designers are currently publishing candidate scores online, but that too seems unethical, and also I do not know if that published data is trustworthy and am hesitant to use it.

For information, a few test creators have sent me their complete data for a particular test of theirs, including candidate names, and I have scored a test by another author (Bill Bultas) myself in the past, so for those tests I have unbiased data.

Another cause of incorrect norms is megalomania by the test creator. There exist authors who delusionally reckon themselves to be profoundly intelligent, but really have much lower I.Q.'s, typically in the 130s to 140s at most. So when they receive test submissions by people whom they perceive as being at roughly their level of understanding, they feel compelled to give out much too high I.Q. scores, otherwise they would have to admit to themselves they are not really as intelligent as they believe.

A median of I.Q. 145 or higher is unrealistic. The high-range population is roughly the upper segment of the general population, cut off at about I.Q. 130. This is not a perfectly clean cut, but if it were, and for the sake of illustration, the following would be necessarily true: With a clean cut at 131 (98th centile) the median would be 135 (99th centile, so halfway the cut and the top). With a clean cut at 135 (99th centile) the median would be 139 (99.5th centile). A median of 145 (99.87th centile) would imply a clean cut at 142 (99.74). This is not consistent with the known population of high-range candidates; most of them are below 142, or at least I believe the evidence for that is more than sufficient.

My experience is that the median of many high-range scores is almost always between 136 and 141. The fundamental cause of this, I think, is that only from the low to mid-130s onward people are interested in intellectual endeavours like taking difficult tests. Below that, it tapers off steeply. Above that, it tapers shallowly, and that shallow curve reflects the actual distribution of those high I.Q.'s in the general population. And this distribution is apparently such that the median of people wanting to take high-range tests ends up around 136-141. The mode is several points lower than the median, the mean several points higher. The mode probably represents the point from whereon the high-range distribution follows the general population distribution (upward). The mode is, more or less, the cut-off point meant in the previous paragraph.

Jacobsen: The following are questions formulated based on input questions provided by Matthew Scillitani. What is the process of making preliminary norms before submissions have been given for a test?

Cooijmans: If it is a fully new test and no data exists for its contents at all, I estimate the minimum raw scores that a Glia Society member respectively a Giga Society should obtain. So for each problem, I look at it and ask myself, "Should a Glia/Giga Society member be able to solve this?" Then I interpolate between those two scores, and extrapolate outward until I reach the edges of the test, where I taper with 5 protonorm points per raw score point. The edges are each sized half the square root of the total possible raw score range.

Jacobsen: There seems to be a stigma around high-range tests. Is there a process to normalize taking them or having them exist in the first place?

Cooijmans: There are indeed many who do not take high-range tests seriously, and this includes prominent figures like the late Hans Eysenck. In one of his "test yourself" books, I remember he was skeptical about the possibility of measuring intelligence in the high range, and even ridiculed it. He provided a number of absurdly complex problems "for the super-intelligent", which appeared to be a parody on high-range testing.

Much of the distrust and denial regarding high-range testing stems from the fear that one might not oneself belong to the most intelligent; it is comfortably reassuring to say to oneself, "Those tests are just puzzles by amateurs and their scores are meaningless, we can not measure intelligence beyond the 99th centile". It is a way to protect one's delusion that no one is verifiably smarter than oneself.

Another cause of the stigma is the inescapable fact that there are fewer women than men in the high range. This is such a taboo that denying the validity of high-range testing is imperative to the politically correct academic, if only for that reason.

A possible process to normalize high-range testing would be to establish it as a recognized branch of psychology at universities. I suspect this would require that we first reverse the decades-long neo-Marxist occupation of academia and make universities into places of genuine science practised by the most intelligent again. A concrete application of high-range psychometrics would be to devise proper admission procedures for universities to undo the dumbing-down that has taken place there over the past half century. The fact that the old Scholastic Aptitude Test and Graduate Record Examination were about the only mainstream tests with validity in the high range illustrates how appropriate high-range testing is in the context of college and university.

For completeness, it should be mentioned that psychologist Lewis Terman (1877-1955) has tried to measure intelligence in the high range with two forms of his "Concept Mastery Test". These were applied to subjects selected as children based on childhood scores of 140 and higher, and followed up in adulthood with the two Concept Mastery Tests. These were verbal tests highly loaded on vocabulary, not permitting references aids. In an unsupervised situation (which was and is how they are typically administered) it is exceedingly easy to cheat on such a test by using dictionaries and thus score absurdly far above one's real level. Also, non-natives of the English

language have a large disadvantage, in the order of 30 I.Q. points. So while these tests were non-robust against cheating and strongly culture-dependent, at least he tried. Since Terman has also been criticized for his belief in eugenics, heredity of intelligence, and racial differences therein, he forms an intersection between high-range psychometricians and hereditarians, so to speak.

Having mentioned the Concept Mastery Tests, I should warn that the scores mostly quoted for them are raw scores, not I.Q.'s. Ronald K. Hoeflin has administered those tests for a while too, also unsupervised, so one should not rely too much on possible reported Concept Mastery scores from test candidates as they may be hugely inflated through fraud.

Jacobsen: Have test construction and norming processes evolved in the aggregate for you?

Cooijmans: Of course, when one has been doing something for decades, one has implemented improvements. If I have to give examples, I have become more concerned with locking in a unique answer and avoiding ambiguity and subjectivity in scoring, and am also inclining more to having tests contain a surplus of difficult problems and a minority of easier ones. Regarding norming, one of the first things I learnt was that z-score equation – equating means and standard deviations – results in incorrect norms because raw test scores tend not to behave linearly, which is required for z-score equation to make sense. So I went with rank equation. Over the years I automated ever more of the process, so that now I can norm a test in 10 to 30 minutes mostly, while originally this took several whole days.

I also learnt to formulate problems better to avoid misunderstanding. For instance, people skilled at mathematics may have a bizarre deformation that makes them interpret numbers differently from normal humans. If I say, "There are three apples on the table", any sane person will understand that there are three apples on the table. But not mathematicians! The mathematician will understand that there are three OR MORE apples on the table. Because the mathematician thinks, "If there are four or five or six… apples on the table, there are three apples on the table too". So to the mathematician you have say, "There are exactly three apples on the table".

Jacobsen: What are the easiest and hardest parts of norming and constructing of a test?

Cooijmans: Easiest: Finishing off the eventual test once the problems have been conceived, and creating the database fields that will receive the incoming submissions. Also, norming is easy on the whole. Hardest: Creating the problems. This has got ever harder, the more tests I made. I try not to repeat myself too much, and try to take into account that the Internet as a search tool has become ever more powerful. The various types of fraud are hard to deal with. I have no sympathy or tolerance for the individuals behind it. The hardest nowadays is to create test problems that are robust against the developments that enable dishonest people to cheat. Those who have spread test answers should reveal the names of the recipients of the answers, so that we can clean up the statistics. And if they sold answers for money, they have to refund, and possible profit they made by investing the fraudulently acquired money should be donated to a good cause.

Jacobsen: Of your tests–51 in-use & 57 retired, which ones are special to you?

Cooijmans: To name a few, Test of the Beheaded Man, Cooijmans Intelligence Test (any form), Daedalus Test, The Nemesis Test, Test For Genius (any form), Only Idiots, The Gate, The Piper's

Test, Dicing with death, The Smell Test. Each in their own way, they demand the candidate to operate at the summit of cognition in ways that are not trivial but tie in to the essence of existence itself. That is what I have generally striven for.

Jacobsen: In pre-2000, you wrote some articles in Netherlandic on test design. Are there any insights from those articles not replicated here or elsewhere worth replicating, or reiterating, here?

Cooijmans: I looked through the articles, and the following points may be worth mentioning:

Marilyn vos Savant occurs briefly in one article; she is known for having "the world's highest I.Q." according to the Guinness book of world records. I would like to add here that someone once showed me a copy of a page from Megarian No. 6 (October 1982) where her actual scores on the Stanford-Binet and preliminary Mega tests are reported. "Megarian" was the journal of the Mega Society then.

Also nice is the early history of Mensa, as related by founder Victor Serebriakoff in one of his books, which was reviewed by David Gamon in the Mensa International Journal of January/ February 1995. The founders at the time believed to be selecting members at the level of 1 in 3000 (some sources say 1 in 6000) but later discovered a mistake in the procedure, as a result of which they had been selecting at 1 in 50. Not wanting to send the bulk of members away again, they left it as it was.

Also mentioned somewhere is Kevin Langdon, creator of the Langdon Adult Intelligence Test (1977 I think) and founder of the Four Sigma society. If one is interested in high-range psychometrics, the statistical reports published by Langdon in the 1970s and later are worth looking at. Langdon's approach differed from Hoeflin's in that Langdon first expressed the candidate's performance as "scaled score" (some conversion of the raw performance) and then equated means and mean deviations of scaled scores and scores on other tests, resulting in a linear relation between I.Q. and scaled score. Hoeflin, on the other hand, normed raw scores directly via rank equation, resulting in a non-linear relation that reveals the non-linear nature of simple raw scores.

This is a good time to explain there are different ways to arrive at a scaled score: The simplest way is to scale raw scores linearly from 0 to 100 or 0 to 1000, for instance. Some test constructors have done that (Alan Aax and Rijk Griffioen, I remember) but it brings no advantage compared to raw scores; the non-linearity of raw scores remains, obviously, when the relation between raw and scaled scores is linear.

If the goal is to obtain a more linear (intervallic) scale, there has to be some weighting or balancing, and a crude but solid method is to give a certain class of problems that appear harder or more important extra credit a priori, regardless of item statistics. This was done by Hoeflin with the Ultra Test, where non-verbal problems get two points. This is effective and without problems, but the resulting weighted scores are still far from linear, if one had any concerns about that.

A more refined way is to give items individual weights based on item analysis. In theory this should yield an intervallic scale, but there are serious disadvantages: (1) A small number of

problems tend to carry most of the weight after weighting thus, which is always dangerous; (2) It adds an extra layer of sampling error because one relies on the correctness of the item statistics, and my experience is that item statistics are not constant but differ from sample to sample, so that one is building on quicksand as it were; (3) The intuitive simplicity of a raw score is lost; the candidate can not know the number of correct answers from the weighted score.

My preference is to use a simple raw score, or in cases where it seems appropriate a crude weighting that does not rely on item statistics, such as in the example of the Ultra Test. If these methods do not result in a meaningful ranking of candidates, that test is bad to begin with and no advanced item weighting will fix it. I accept that raw scores are non-linear, and the conversion to linearity takes place in the norming of raw score to I.Q.

That last sentence leads to the question, "How do we know that I.Q. is a linear scale?" The answer is that I.Q.'s are deviation scores; they denote a distance to the mean in a hypothetical normal distribution. Note the word "hypothetical"; it is not claimed that intelligence follows a normal distribution in the physical reality. But the tacit assumption in statistics is that when a distribution is normal (Gaussian), its underlying scale is linear (intervallic). So when you force test scores into a normal distribution, you create a linear scale, or that is the unspoken idea. This is expressed in the way we identify points on the scale in terms such as "2 standard deviations above the mean". This implies an underlying linear scale; after all, if the scale were not linear, the one standard deviation would not be the same as the other, so it would make no sense to say "2 standard deviations above the mean"! In fact, the mere computation of an arithmetic mean assumes an underlying intervallic scale, as it involves summation.

So the bottom line is, if we take care that the frequencies of I.Q.'s beyond various points of the scale do not differ too much from their theoretical rarities in the normal distribution, we may assume that I.Q. is linear. I say this without claiming that deviation I.Q.'s are the best way to express intelligence; but I do not have a better way at the moment.

Jacobsen: Some submitted questions anonymously. These are the adaptations of those questions: Personally, do you know any geniuses? If you do not know any personally, where are all of the geniuses?

Cooijmans: I have to say that when it gets anonymous, the quality goes down. Imagine that I answered "no" to the first question! How insulting that would be to everyone I know! Since a genius is someone who exercises a lasting influence in any field, inherently it can only be known in hindsight who was one, like long after the genius' death. It is well possible that several people I know will turn out to be geniuses, but we do not know yet who they are.

In history books you will find a lot of identified geniuses.

Jacobsen: Why refer to these individuals in this way, i.e., as geniuses? What traits characterise them?

Cooijmans: The word "genius" comes from the Latin "gignere", meaning to conceive, to bring forth, to cause. Francis Galton used the word "eminence" for what is now mostly called genius.

The traits of genius, according to me, are intelligence, conscientiousness, and a wide associative horizon. Genius is not talent. It requires talent, but talent alone does not suffice. One will need to apply that talent in order to make a lasting contribution.

Jacobsen: Do you see yourself as a genius? If so, why? If not, why not?

Cooijmans: Naturally, someone of my enormous modesty and humility would never call oneself a genius. I leave that to the scores of future generations who will devote their lives to the study of my work.

Jacobsen: What do you think has been the contribution of your I.Q. Tests for the High Range? Is it a work for study by others or a hobby?

Cooijmans: The contribution lies in studying the measureability of intelligence in the high range, and some other questions related to that as stated at https://iq-tests-for-the-high-range.com/mission.html . It is certainly worthy of being studied by others, and others should also undertake such study independently. It is not just a hobby, except in the sense that one can make one's hobby into one's work.

Jacobsen: Who are others who you see like yourself in studying high ranges of intelligence?

Cooijmans: This can only be answered properly for people who were (already) working longer ago, before the current generation of high-range testers. That would be Lewis Terman, Kevin Langdon, Ronald K. Hoeflin, Xavier Jouve, and Laurent Dubois. For the ones who came after these, it is too soon to judge their merit.

In addition, there have been some people who created tests that looked truly good to me, but who only kept scoring their tests briefly and then withdrew from testing, so that little or no usable data resulted. These people exemplify what I said a few questions ago: that talent is not genius, but merely a requirement for genius. They had talent, but did not use it to make a lasting contribution to high-range testing.

Jacobsen: What is the most common mistake people make when submitting feedback about your tests?

Cooijmans: Assuming that they have understood a test item correctly, and then commenting on it from that assumption.

Jacobsen: What aspects of people's test feedback seem confusing?

Cooijmans: It can be confusing if people send feedback before sending answers. I have to be careful not to help them by responding. Nevertheless, in case the feedback concerns a mistake in a test problem, it can be useful, especially when a test is very new.

Jacobsen: The most common Marathon Test Numeric Section score is a perfect 44 out of 44. What lessons have you learned from this high-end score saturation?

Cooijmans: That the problems are not hard enough. Also that a series of similar problems of increasing difficulty tends to be too easy on the whole. And that, to make a numerical test hard enough, either very difficult mathematics-biased problems are needed, or problems that

implement a pattern that needs to recognized. The latter seem the most fair, the former seem to give an advantage to people skilled at mathematics.

Jacobsen: When creating high-range questions, is there a consideration of steering test takers toward wrong answers? Are extant questions ever modified in this way?

Cooijmans: Obviously, steering test-takers toward wrong answers is the whole point of creating good test items, not only in high-range testing but also in mainstream intelligence tests. There is even a word for it: distractors. Multiple-choice tests, omnipresent in mainstream psychological testing, contain answering options that are wrong but appear more plausible than the intended correct answer.

Thus, a candidate who really can not solve any problem at all will score below the chance guessing level, and this lower level is called the "pseudo-chance level". For instance, if a test has 40 problems and 5 answering options per item, the chance guessing level is 8 correct, but the pseudo-chance level may be only 5 correct due to distractors.

Extant questions are not generally modified in this way.

Jacobsen: Which books or literature, even individual articles or academic papers, on psychometrics have provided helpful accurate understandings of psychological measurement, psychometric concepts, etc., for you? Others may find some fruitful plumbing there.

Cooijmans: The most specific sources regarding high-range testing are the various statistical or norming reports by Kevin Langdon and Ronald K. Hoeflin, as issued by them in the 1970s through 1990s (Hoeflin only started in the 1980s I think). These helped to see how high-range tests are normed, and also aided in the interpretation of scores on a lot of those old American tests like the Scholastic Aptitude Test, Miller Analogies Test, Army General Classification Test and more. The "Omni sample" of the Mega Test contains many scores of those versus Mega Test scores, and as such is an important anchor of high-range tests to the general population, especially so since those old tests in some cases did discriminate into the high range. This can not be said about the newer dumbed-down versions of the educational and military tests, whose validity tends to end at the 99th centile, and for whose interpretation one should consult the information provided by the relevant issuing organizations.

What one can see for instance in the Omni sample is that the old G.R.E., S.A.T., and Army General Classification Test correlated quite well with the Mega Test, while on the other hand the Wechsler Adult Intelligence Scales and Stanford-Binet, by many regarded as the gold standard of I.Q. testing, appeared to lack any validity in the high range. This observation holds true until today in data collected by myself, except for the old Army General Classification Test on which I have almost no data.

Then, an actual text book on psychometrics I have studied is the Netherlandic "Testtheorie" by P.J.D. Drenth and K. Sijtsma from 1990. This covers both classical psychometrics and the newer item-response theory.

Another useful book, a bit more general, is "Applied statistics for the behavioral sciences", second edition, by Hinkle, Wiersma, and Jurs, from 1988.

An important book on intelligence testing is “The g factor” by Arthur Jensen, from 1998. While not intended as a psychometrics text book, it does contain a lot of advanced information on psychometrics, including some factor analysis, often in the footnotes.

As it happens, there is also an e-book called “The g factor” by Christopher Brand, from 1996, also containing information on psychometrics and some factor analysis.

A book on statistics in general (not psychometrics) I have studied is the Netherlandic “Statistiek in de praktijk” by David S. Moore and George P. McCabe. I see there is an English version too, “Introduction to the practice of statistics; 2nd edition” (1993).

A book dealing specifically with multivariate statistics such as correlation, regression, and factor analysis is “Using multivariate statistics” (third edition) by Barbara G. Tabachnick and Linda S. Fidell (1996).

I also still have my mathematics books from secondary school, one of which contains chapters on statistics and probability calculation. Occasionally I look through those to refresh these basics of my knowledge in this field.

Finally I want to add that the history of statistics and of mathematics is informative regarding psychometrics. Reading about such will teach you that statistics has been closely related to psychological testing since the 19th century, and that probability calculation was developed for the purposes of gambling and insurance.

The history of mathematics in general, found for instance in “A concise history of mathematics” by Dirk Jan Struik, tells us that mathematics originates in the early days of agriculture, cities, and large-scale administration. That is, within the past ten thousand years or so, the holocene, after the last glacial period. Computing the area of parcels of land required mathematics.

I suspect that the intelligence of the people coming out of the last glacial period was primarily of a visual-spatial nature, and as they became settled and practised agriculture, built cities, and administrated societies, they needed higher numerical ability as well as written language. I imagine that spoken language existed long before that, originally in the form of words without grammar some two million years ago to coordinate hunting in groups in early Homo, and later on with grammar, perhaps in the days of Homo sapiens.

Language is not unique to humans incidentally, but exists in other beings too, such as birds, primates, and whales. Animals like crows are likely at the intelligence level of early Homo, but I am uncertain if their physicality will allow a further development such as has taken place in Homo. Key points like the manufacturing of tools and mastery of the fire may require arms, hands, fingers, and thumbs such as humans have.

Visual-spatial ability is also not restricted to humans, but found in many animal species, in particular to enable predating. As such, visual-spatial ability should be a few hundred million years old, as that is when the first predators came.

The importance of this history of abilities is that when we test abilities now, the results we get, such as the intercorrelations of various abilities, are as it were a fossil record of this evolution. A development that I believe takes place in civilized societies is the erosion of the original visual-

spatial ability in favour of verbal ability. A high level of verbal ability in the absence of the foundation of visual-spatial ability, I think, leads to dishonesty, deceit, evil, decadence, and societal collapse.

Jacobsen: Thank you for the opportunity and your time, Paul.

Cooijmans: I never know what to respond to here.

Alex Tolio

2024-11-22

Alex Tolio is someone interested in I.Q. tests and high-range test construction. He discussed common errors in creating valid, reliable high-range IQ tests, highlighting test constructors' biases, insufficient abstraction, and questions overly reliant on novelty or complexity. Tolio notes that a balance between creativity and clarity is essential to prevent ambiguous items. He suggests beta-testing, revision, and removing unnecessary elements to reduce bias. He advocates for heterogeneous tests to best capture the general intelligence factor (g) and highlights the importance of correlating new tests with validated ones. Confirmatory factor analysis and careful data preparation are key for precise measurement. Feedback issues often involve claims of ambiguous items.

Scott Douglas Jacobsen: What are common mistakes in trying to make high-range tests valid, reliable, and robust?

Alex Tolio: I suppose the most common mistake may be one's own biases towards what they consider valid.

While there exists objectivity in what parameters constitute a good test, there does not seem to be a set consensus on what really is "absolutely needed" to measure IQ's supposedly above 3-4SD.

I can only suppose with certainty that the need for increased abstraction (rather than speed) has to be present, but there exists a fine line between what is really abstract or simply convoluted.

Because of this, test-creators (including myself), are likely implementing their own bias within the question(s) themselves, which ultimately may reflect one's own attitude to measuring very high IQ. In other words, what one considers great items or perhaps even format, is already biased.

I believe this can manifest itself in many forms; such as:

1. A tendency towards creating items that may be inherently polluted or prone to by external factors; examples of such may be unnecessary layers in hopes of increasing difficulty, which, when made poorly may falsely promote novelty in cost of quality. I believe the main flaw of such items may be that a conscientious but yet less capable individual; may be able to untangle them given enough time. This means that the difficulty of the item in question is shallow; given it may be sufficiently understood with enough time, it is only a mere product of the layers of which the true solution may be.
2. Implementation of too much creativity, I've found this may be counter-intuitive to creating a solid test, but I believe there need be a balance. In the process of one's own journey to create novelty, one will be faced with such issue, where a large risk is esotericism, which may ultimately bring its own set of problems. Such items are also prone to interpretation, which ultimately hurts a test's quality, because the reliability is primarily based on responses given by the candidates that answered an item. Because reliability is a product of this, such questions will dramatically lower the consistency, because of the variety of the responses received. I believe there are numerous tests out there of which this is an issue; and possibly a reason of the lack of report for such statistic.

To note, a test may be good in of itself, but the sample to which it is administered will have a direct effect on its quality.

3. Misjudge of sufficiently discriminating questions may result a test author to produce a test that is too difficult.
4. Some test authors are not attempting to create tests to measure IQ, but is rather a showcase of their own ability, in which superfluous and unreasonably difficult items are present.
5. Recycling one's own questions (something I'm equally guilty of)!
6. Poor balance of difficulty within a test; or too "graded" difficulty. If difficulty is too calculated, or a steep sudden increase occurs, there may exist a common "wall" that collapses in a certain raw; of which people of a ball-park of ability may fall into, possibly seizing to discriminate further.

A good example of that is when a norm has a jump of almost +10-15 points per raw; this means that the people who are in this raw are not distributed.

7. Creating too many difficult questions, does not make for a great test as a whole.

8. Possibly the repetitive use of patterns, of which one has seen, thought of, or previously used in tests. This creates a author-specific learning effect, of which familiarity with such pattern(s), assuming they have been solved once, will not require sufficient amount of ability afterwards.

I believe this is also a mistake I tend to make. However, I believe it is ultimately unavoidable as one proceeds to solve an author(s) tests. One will have to exert the most effort in the first couple tests, of which may truly measure "IQ". But as one proceeds to take multiple tests by an author, their scores are prone to increase. I truly think this reflects that tests may seize to be robust in particular cases; this seems to be irrespective of whether author's work is of quality, but may be related to becoming more aware of what may be asked by a certain author; developing an intuition for such; or perhaps an eventual reverse-engineer on how the author may tend to think, and general familiarity. This, for me, highlights the human error that is present in such tests, which constitutes imperfect measurement; but also possibly invalid after a certain period of time.

9. Test questions do not discriminate properly, as-per-flawed construction. This ties well with shallow questions. I believe questions that poorly discriminate are likely the ones that are also prone to be more polluted by (external) factors required; that is in addition to pattern recognition (core).

Whereas those factors may ultimately eliminate the quality and need (albeit not absolute) for pattern recognition.

10. Items that are poor may result in an "artificial ceiling" of which the test does not bare the ability to truly measure to the levels it is reporting.
11. False or manipulated statistics (or lack thereof)

Jacobsen: What are the core abilities measured at the higher ranges of intelligence or as one attempts to measure in the high-range of ability?

Tolio: I believe what is unitary of such tests is their increased demand on abstraction.

This means that the focus is shifted to the depth rather than speed, of which may challenge the strength of ones understanding.

Because understanding requires pattern recognition as per item, candidates answer directly reflects their understanding of an item.

Assuming answer is commonly unique, then the discrimination should occur naturally; a phenomenon better seen by the (likely) common understanding of items between candidates of different ability levels.

This does not necessarily imply there truly is an alternative (or weaker answer), but is a consequence of lack of sufficient understanding, of which ambiguity may seem apparent.

However, it is also possible there does exist a weaker "common", of which ideally should be eliminated.

The strictness of the answer is perhaps also a prerequisite, of which candidates of higher ability tend to be extremely rigorous, and may not rush to think they have "the" answer.

In short, the core abilities are depth of understanding (reason), pattern recognition, higher ability for abstraction, and often inherent divergence that may result.

Pattern recognition of the highest form, in my opinion, is not about recognizing a pattern in all the chaos (or noise), but proposing structure to it; such that it is sufficiently and simply understood.

This means unnecessary elements implemented in questions may contradict this notion, and presentation ought to be clean.

If an element does not bare meaning, or is not useful in any particular way, is probably best removed. Which may also reduce item likelihood to load on external polluting factors.

Jacobsen: How do you remove or minimize test constructor bias from tests?

Tolio: It is not possible in its absolute form, but certain measures may help to eliminate it.

-Questions should (ideally) be beta-tested before release.

-Conscious effort of elimination by author, rigorous revision of items.

-Creating questions that are more pure; so that cultural difference is minimized

-Exclusion of esotericism and the inherent creativity

-Removal of unnecessary elements, promoting a clean presentation of an item

-Too many clues may be as good as none

-Very careful implementation of novelty, and preferring a more universally understood form of expression.

-Eliminating the projection of one's own subjective ideas of sufficient discrimination between higher-ability candidates;

Jacobsen: What should be done with homogeneous and heterogeneous tests?

Tolio: Despite my own construction of homogeneous tests, it is clear that heterogeneous tests are the best measure.

It is not possible to extract true g if the test is homogeneous, and is likely extracting the most "general" factor the test is loading on.

I believe heterogeneous tests are the only way to extract g, because it is also simply a manifestation of a unitary excellence in various cognitive facets;

I've come to the recent conclusion that it is best that a test's ceiling should be raised by this variety, rather than the implementation of extremely difficult questions in one area.

The latter being flawed, as these questions are the most challenging to make "healthy".

1. Cooijmans tests are a great model, but I would also encourage possible "inventions" for authors. Such forms may be of the problems or task in question.

Jacobsen: What tests and test constructors have you considered good?

Tolio: P. Cooijmans has the most thorough and professional work, which could additionally be used for educative purposes.

1. Jouve has professional work; and implements a rather linear approach. Rather accurate; however I believe it could be potentially limiting after an arbitrary threshold.

And for this I think Johnathan Wai is a better fit.

Johnathan Wai's work is more of well-suit to the high range.

I am not very well acquainted with old generation of authors, or non-western.

1. Prousalis work is promising.

There are other authors of which I like, but may lack substantial statistics.

Jacobsen: When trying to develop questions capable of tapping a deeper reservoir of general cognitive ability, what is important for verbal, numeral, spatial, logical (and other) types of questions?

Tolio: While overlaps with the previous, however I will add:

Question difficulty may be effectively raised by:

Careful implementation of additional rules and the complexity of them; (reason)

Increasing the abstraction level of the idea in question; (horizon)

These seem to be primary ways, but not necessarily only ways.

In the process of a candidate tackling an item, one must first generate a plethora of ideas of whom (most, may not be meaningful) in solving the item.

However; the discrimination of such items occurs as per their idea, and not necessarily that they are difficult to work-around (reason).

It is likely that if one is not capable, they will never generate the idea to solve the item in question, of which requires associative horizon (see P. Cooijmans).

The most important part of every item is whether discrimination occurs; while it may seem intuitive that an author should be very rigorous with their own idea(s) and logic, or to even the disambiguation of their own items; this perfection is not always necessary in the process of creating a good item.

I've come to observe that an author mainly needs to think of the best application of their idea, in combination to the best presentation of such; of which, if made correctly,

should naturally dis-encourage alternative solutions from being present. In other words, disambiguation may occur as a consequence of this.

Jacobsen: What is efficient means by which to ballpark the general factor loading of a high-range test?

Tolio: I suppose since ballpark is used:

Correlation of authors test to a test or test(s) known to have captured g.

This means that correlations with professional tests are absolutely necessary, because they indicate construct validity.

Jacobsen: What is the most precise or comprehensive method to measure the general factor loading of a high-range test, a superset of tests, or a subset of such a superset?

Tolio: The most precise method I believe would be confirmatory factor analysis, of which samples are usually not apt.

This means it may misrepresent a test's true g loading, and this is not necessarily towards the "better", but often contrary, however it can be both ways.

It is also very likely that the extracted factor is not true g, of which a test may have extracted A "general factor".

Correlations help to ensure of this.

When omega η is estimated, the square root of omega η is thought to be the g loading.

Data may need be prepared very carefully, but meaningful factor analysis may be least possible with a clean sample of N=70-100, albeit not precise.

Intercorrelation between tests, or one's own tests, may be flawed because of the assumption the tests have captured g; unless correlations previously made have indicated construct validity, of which this method may be then considered valid.

Jacobsen: Have test construction and norming processes evolved in the aggregate for you?

Tolio: It started as a creative outlet and hobby for me; and I believe such lack of seriousness may be vaguely reflected, however, I attempt to provide honest statistical work/methods. It has been (and still is) a learning curve for me, of which I've tried to educate myself from the variety of work provided by authors, particularly P. Cooijmans, and using it as a stepping-stone to my own conclusion(s). There is still a lot to learn, in fact I would claim that I'm not even remotely advanced with statistics.

So I may use a disclaimer in the case something said is of ignorance; P. Cooijmans norming method is the definite best for norming a high-range test. Because of the linear issue proposed by Z-scores.

I have tried to propose at the very least my own, of which seems to have began with a flawed premise.

I proposed that many high range IQ tests do not contain sufficiently discriminating items; such promoting an essentially "artificial ceiling".

This means that the items may not scale further upon a certain level, and are likely reporting false IQ scores beyond that.

I thought of attaching a "grade" or an IQ level for each problem, of which may be estimated by its solvability, of which is tied to the reported IQ of candidates that solved this item,

of which try to extract the least "known" IQ possible to solve a question, thus attaching this grade to the question.

The reason of this, would be to ball-park the true IQ content/difficulty present within a test.

This would help with promoting a healthier ceiling, or at the least; estimating true discrimination of a test.

Some but not all flaws:

The flawed premise is that a question absolutely necessarily needs a minimum IQ, this assumes that every question is good.

And of course; the dishonest reports of IQ's, or scores.

Jacobsen: What is the most common mistake people make when submitting feedback about your tests?

Tolio: I've received primarily positive feedback, however there are often cases of which candidates attempt to point out a supposedly ambiguous item.

To eliminate this issue, whilst being entirely aware of the feedback this gives; I report every single "incorrect" answer to items.

I hope this approach is transparent enough, so that there should be no further discussions.

I'm fully aware however that such approach (and availability of tests), is promoting further familiarity/practice.

I hope to tackle this issue by creating more novel items.

However, it has also served as an experiment for me, and thus far it does not seem to be of extreme benefit to candidates.

This means that thus far; despite reporting the incorrect answers, the correct answers are still not "correctly" found, by re-occuring candidates in case an item may have been re-used.

This does not necessarily prove the contrary about the approach, but it may be of lesser impact as initially thought of.

Patrick Liljegren

2024-11-22

Patrick Liljegren is a member of the Synaptiq Society and The Glia Society. He achieved a 171 on the IQ test OASIS. Patrick is passionate about audio system tweaks, including the use of crystals. He enjoys eating banana ice cream daily and braving the winter without a jacket. His philosophy emphasizes self-improvement, perseverance, and the importance of having fun. Liljegren's passion for test construction developed from his interest in cognitive processes and personal experience with traditional IQ tests, which felt limited in measuring deeper logical reasoning. Dissatisfied with repetitive and uninspiring test formats, Liljegren sought to create engaging and enjoyable tests that foster cognitive growth and reflect true intellectual ability. He emphasizes the importance of avoiding biases, maintaining rigorous test design, and ensuring test reliability. His focus is on holistic, multi-domain questions that stimulate deeper problem-solving. For valid results, he values participant engagement and careful test scoring, addressing potential errors and aiming to support test-takers' growth and confidence.

Scott Douglas Jacobsen: When did this interest in test construction truly come forward for you?

Patrick Liljegren: This interest in test construction emerged as a natural evolution of my lifelong passion for understanding cognitive processes. After taking numerous IQ tests myself, I realized there was a distinct kind of logic I was using—one that allowed me to engage with problems on a deeper level than these tests seemed to measure. Each test left me with the sense that there was something more to explore, a way of thinking that wasn't captured by standard approaches.

Over time, I grew increasingly curious about crafting an alternative that included this "deeper logic." I wanted to create tests that not only challenge conventional boundaries but also engage test-takers with a format that's more enjoyable and dynamic than rows of text or numbers. In combining rigor with creativity, I hoped to produce tests that would captivate people who, like myself, crave an experience that reflects the true intricacies of high-range cognition. This blend of challenge and engagement was a calling that I couldn't ignore, and that's when I truly began dedicating myself to test construction.

Jacobsen: At the time, what were the realizations about the tests and the need to develop yours?

Liljegren: My biggest realization about other people's IQ tests was that they were, frankly, very boring for most people. The majority of the tests I encountered were filled with repetitive tasks—just text and numbers—making the experience feel more like a tedious obligation than an engaging intellectual challenge. It became clear to me that this format didn't capture the interest of participants, and I could see why many would be unwilling to invest hours on end in something so monotonous. This realization drove me to create my own tests, ones that would not only be intellectually stimulating but also fun, encouraging people to engage deeply with the process without feeling like it was a chore.

Jacobsen: What are common mistakes in trying to make high-range tests valid, reliable, and robust?

Liljegren: A common mistake in creating high-range IQ tests is neglecting the participant's experience, which can lead to a lack of engagement. If the test is so boring or tedious that people don't want to spend the necessary time on it, then the test fails to be truly valid. Cognitive ability cannot be accurately measured if participants are rushing through questions or abandoning the test early. A valid test needs to capture the participant's full cognitive range, which means making it not only intellectually challenging but also engaging enough to keep the participant involved. Without this engagement, the test becomes more of a hurdle than a meaningful assessment.

In the past, when I was a teenager, I took an IQ test administered by a psychologist. At the time, I wasn't particularly interested in the test, so I rushed through it, not fully engaging with the questions. The result was that I received a score which they deemed as 'retard level,' and the feedback was devastating. However, this experience was incredibly eye-opening for me. It made me realize how much external factors—like lack of interest or engagement—can skew a test result, leading to an inaccurate and harmful assessment of one's abilities. That experience, while painful, inspired me to create tests that are not only challenging but also engaging enough to allow people to truly demonstrate their cognitive potential, free from the constraints of traditional, often discouraging, testing methods.

Jacobsen: What are the counterintuitive aspects in taking tests and making tests in the high-range?

Liljegren: A major issue with traditional tests is that people are often more focused on completing them as quickly as possible to achieve a higher IQ score. They rush through the questions without taking the time to dive deeper into the problems, which leads to surface-level

answers. This tendency to prioritize speed over depth contributes to the overall boredom with text- and number-based tests, preventing individuals from fully exploring the underlying logic or the deeper meanings within the questions.

In contrast, my tests break away from this model by incorporating images and humor, which serve to keep participants engaged and entertained. This approach fosters a sense of enjoyment and curiosity, encouraging participants to delve into the problem-solving process with a genuine desire to explore deeper, rather than just rushing to finish. The inclusion of humor makes the experience more relatable and less daunting, while still maintaining intellectual rigor, creating a more enjoyable and effective testing experience.

Jacobsen: What are the core abilities measured at the higher ranges of intelligence or as one attempts to measure in the high-range of ability?

Liljegren: The core abilities associated with high-range intelligence are not about speed, as calculators and computers are not truly intelligent. Real intelligence involves the ability to approach a problem from multiple perspectives and select the most logical solution. It's about having a broad understanding of various concepts and identifying the most reasonable choice. Rushing to solve something often limits the range of perspectives considered. True intelligence requires taking the time to explore different ways of solving the same problem, which can take thousands of hours. It's akin to exploring a forest and visiting every part to fully understand it and create a complete mental map. This process of deep exploration is essential for making the best decision.

Jacobsen: In an overview, what skills and considerations seem important for both the construction of test questions and making an effective schema for them?

Liljegren: If the author of a test has certain mental limitations or a narrow understanding of basic human behaviors, it can lead to a biased, limited test. This test would reflect the author's own cognitive framework, potentially making it skewed toward a specific way of thinking. If such a test is used, individuals who score highly may be seen as having similar mental characteristics or limitations, which could simply be a result of the test's narrow perspective.

In this case, the high scores would no longer represent general intelligence or cognitive flexibility, but rather a shared bias or limitation that the test fails to account for. This would be problematic, as it would reinforce the cognitive limitations of both the author and the test-takers, rather than providing a comprehensive measure of intelligence.

For a test to be valid and accurate, it must be free from these personal biases, ensuring it measures general intelligence, not a particular mindset or cognitive limitation.

Jacobsen: Any thoughts on proposals for dynamic or adaptive tests rather than–let's call them–"static" tests consisting of a single item or set of items presented as a whole test, unchanging, instead of a collection of algorithmically variant or shifting items adapting to prior testee answers in a computer interface?

Liljegren: I believe the most effective dynamic test would be a virtual world where time is infinite, and the participant cannot escape until they have successfully solved the test. In this

environment, the test taker's intelligence would be measured based on the choices they make and the paths they explore as they interact with the world around them.

The absence of time pressure is crucial. Just like in real life, there's no strict timeline for decision-making, and rushing would only limit the depth of exploration. Early choices might lead the participant down a seemingly wrong path, but without a time limit, they would have the opportunity to revisit earlier decisions, re-evaluate their choices, and learn from their mistakes. This reflects the process of growing through experience and finding the best path through reflection and exploration.

The world would be locked, meaning the participant cannot escape or finish the test until they have solved it in their own way. This would ensure that the participant is fully engaged and has the space to explore every facet of the test, allowing for a deeper understanding of their own problem-solving process and decision-making abilities. The test is not a race—it is a journey where intelligence is reflected in adaptation, learning, and growth over time.

By removing time constraints and providing an infinite amount of space to explore, this dynamic test would truly measure the ability to think critically, adapt to new information, and learn from past choices. The best answers would come not from rushing but from the thoughtful, reflective process of solving problems over time in an ever-evolving environment.

Jacobsen: How do you remove or minimize test constructor bias from tests?

Liljegren: I believe that to minimize bias in test construction, the author needs to possess a high level of general intelligence. The higher the intelligence, the broader and more flexible the thinking, which helps in considering multiple perspectives and reducing narrow-minded bias. Lower intelligence tends to create more biased and limited thinking, which may not resonate with a diverse range of test-takers.

Moreover, the test constructor needs to have a well-rounded understanding of human behavior. A test creator with greater intelligence is more likely to recognize these biases and account for them, ensuring that the test is fair and representative of a wider audience.

Jacobsen: How do we know with confidence many listed norms are, in fact, reasonably accurate on many of these tests? What is the range of sample sizes on the tests, even approximately, now? Practically speaking, for good statistics, what is your ideal number of test-takers? You can't say, "8,128,000,000."

Liljegren: I believe the sample size for a test should consist only of individuals who take it 100% seriously. When test-takers are fully engaged and committed, the data collected will be far more accurate and reliable. This ensures that the results reflect the true cognitive abilities of the participants, rather than being skewed by rushed or careless answers.

While a larger sample size can be beneficial for diversity and generalization, the quality of the responses is paramount. A smaller, but more focused group of serious participants will yield more valid and meaningful norms. In essence, the test becomes much more accurate when the sample is composed of individuals who approach it with the same level of seriousness and dedication as athletes competing for a gold medal.

In my opinion, a sample size of around 100 highly engaged participants is ideal for creating accurate and reliable test norms. When the group is larger, such as 500 participants, the level of seriousness tends to decrease, especially for those who find themselves near the bottom of the results. As a result, these individuals may rush through the test or fail to fully engage, which can lead to less reliable data.

In contrast, when the sample size is smaller, like 100 participants, the competition feels more real and the stakes are higher. This creates an environment where test-takers are more likely to commit fully to the process and give their best effort. The focus and dedication of such a group result in more meaningful and precise data, as each participant is genuinely invested in the outcome, much like athletes competing for a gold medal. By keeping the sample size smaller and more engaged, the test is able to capture more accurate measures of intelligence and cognitive ability.

Jacobsen: Is English-based bias a prominent problem throughout tests? Could this be limiting the global spread of possible test-takers of these tests rather than limiting them to particular language spheres? Although, these tests are taken, to a limited degree, in many countries of the world in all/most regions of the world.

Liljegren: I avoid English-based bias in my tests by incorporating numbers and images, which are universal elements that don't rely on language. This approach ensures that both English and non-English speakers have a fair chance to perform based on their cognitive abilities, rather than language proficiency. While English language knowledge can be an advantage for native speakers, non-native speakers might actually have an advantage in some cases. When they encounter unfamiliar words, they are more likely to look them up, which could reveal subtle clues that a native speaker might overlook, assuming they already know the meaning.

Overall, this approach balances things out and ensures fairness for both groups. By relying on numbers and images, I make sure that the test evaluates true cognitive skills, regardless of the language spoken. This way, the test becomes more universally accessible while still maintaining its reliability across different populations.

Jacobsen: When trying to develop questions capable of tapping a deeper reservoir of general cognitive ability, what is important for verbal, numeral, spatial, logical (and other) types of questions?

Liljegren: I believe that in designing questions that tap into deeper cognitive abilities, it is crucial to integrate different domains—verbal, numerical, spatial, and logical—into a cohesive, interconnected framework. Rather than treating each domain as separate, questions should challenge test-takers to blend these different types of reasoning, creating a more holistic and real-world relevant measure of intelligence.

In traditional tests, we often see sections devoted solely to one type of reasoning: a verbal section, a numerical section, a spatial section, etc. However, true cognitive ability is more complex. It lies in how well someone can synthesize and apply knowledge across these domains to form a broader, unified understanding. This is akin to solving a puzzle where the pieces are of different shapes and forms—verbal, numerical, spatial, and logical. The challenge comes not

from solving each piece individually but from recognizing how they interconnect and contribute to the whole.

This integrated approach mirrors real-world problem-solving, where we constantly draw upon diverse areas of knowledge. To test someone's true cognitive abilities, we must create challenges that require them to blend these elements and think beyond linear, compartmentalized patterns. It's about understanding the bigger picture and making connections that others might miss, which is often the hallmark of high-level cognitive processing.

In this way, the test becomes more than just a measure of isolated skills. It gauges how well someone can think creatively and flexibly, applying various types of reasoning in novel ways to solve complex problems. This method of testing challenges individuals to think out of the box, drawing on multiple domains to find the best possible solution—a far more comprehensive reflection of intelligence than traditional, domain-specific tests.

Jacobsen: What are roadblocks test-takers tend to make in terms of thought processes and assumptions around time commitments on these tests? So, they get artificially low scores on high-range tests.

Liljegren: Many test-takers often fall into the trap of underestimating the time commitment required for high-range tests. They tend to think that they should be able to answer the questions quickly, driven by a sense of confidence or even narcissism. These individuals often assume that their initial answer is correct without fully considering alternative perspectives or exploring the problem deeply. This overconfidence typically leads them to rush through the test, which results in lower scores.

The key to performing well on such tests lies in approaching them with humility. When a person is humble enough to accept that they don't have all the answers and that there could be multiple ways of thinking about a problem, they tend to spend more time reflecting on each question. Rather than sticking rigidly to their first choice, they're open to exploring different avenues and rethinking their responses. This deeper, more methodical approach leads to better performance, as it allows them to tap into a broader range of insights and avoid missing crucial clues that could improve their answers.

So, in essence, those who approach a high-range test with an open mind and a willingness to consider all possibilities—without rushing or prematurely settling on answers—are far more likely to succeed.

Jacobsen: What is the intended age-range for high-range tests? How do these account for individuals younger and older than this range?

Liljegren: I believe that the intended age range for high-range tests often reflects the cognitive and emotional maturity required to fully engage with them. Younger individuals tend to rush through questions and make quick decisions, often due to a lack of experience or an overestimation of their ability to answer immediately. As people grow older, they gain a sense of relaxation and wisdom that allows them to approach problems more thoughtfully. This maturation process helps individuals realize they don't have all the answers right away, which leads them to spend more time considering different perspectives and refining their responses.

When I was younger, I would only spend a couple of hours on each test, but now, after years of experience, I dedicate thousands of hours to fully exploring every test I take. This shift in approach illustrates how cognitive growth and emotional development over time lead to better results on high-range tests.

Jacobsen: What is important in constructing and norming a test?

Liljegren: When constructing and norming a test, one crucial factor that is often overlooked is the cognitive growth that occurs during the testing process. A test that promotes cognitive development as the participant moves through it is not only more engaging but also yields more accurate results. This dynamic approach ensures that the test-taker's cognitive ability is allowed to evolve, which, in turn, enhances the reliability of the results.

Cognitive Growth During the Test:
One of the most important elements in test construction is ensuring that the test encourages growth in cognitive ability while the participant is engaging with it. This process involves crafting questions that require test-takers to think critically, adapt their strategies, and explore new methods of problem-solving as they progress through the test. By introducing progressively more challenging and thought-provoking questions, the test encourages test-takers to evolve their thinking, enhancing their problem-solving ability and, in turn, their cognitive growth.

This improvement is especially important because it directly influences engagement. When a test-taker sees their cognitive abilities growing during the test—when they feel that they are not just answering questions but also becoming more intelligent throughout the process—they are far more likely to invest the necessary time and focus to fully engage with the test. This increased focus and effort can lead to a more accurate and comprehensive assessment of their potential, as they are operating at their maximum cognitive capacity.

Engagement and Accuracy:
As a test-taker becomes more engaged in the process and experiences cognitive growth, they are more likely to take the time to consider their answers carefully and explore multiple perspectives before finalizing them. This is where the real value of cognitive growth comes into play: when participants are learning and improving as they work through the test, their final answers are more reflective of their true cognitive ability. They are less likely to rush through questions, make careless errors, or settle on superficial solutions.

In contrast, tests that are too short, or lack this cognitive growth element, may encourage rushed decision-making, ultimately leading to less accurate results. In such cases, the test may not fully capture the test-taker's potential, and the results could be skewed by the lack of cognitive engagement. Therefore, a test should not only measure raw ability but also stimulate growth throughout its duration. By doing so, test-takers' cognitive abilities are fully exercised and measured at their peak.

The Importance of Test Length:
For this process to take place, the test needs to be long enough to allow for meaningful cognitive growth. If the test is too short, test-takers will not have sufficient time to experience this transformation. As the test progresses, their problem-solving skills improve, which should be

reflected in their answers as they revisit and reconsider earlier questions. This iterative process ensures that their final performance represents a more accurate picture of their cognitive abilities.

By fostering cognitive growth during the test, you are not simply assessing the static intelligence of the participant; you are capturing the dynamic nature of their cognitive abilities. This allows for a much more nuanced and accurate understanding of their intelligence, which is crucial when norming the test. This approach can lead to more meaningful norms, as test-takers are measured based on their full cognitive potential, not just their initial capacity.

In summary, test construction and norming should go beyond merely measuring cognitive ability at a fixed point in time. By designing tests that promote cognitive growth, you engage test-takers in a deeper and more meaningful way, which not only improves their performance but also leads to more accurate, reliable, and comprehensive results. This dynamic approach is essential for creating a test that truly measures the depth and breadth of human intelligence.

Jacobsen: Cheaters exist. Frauds exist. How do you a) deal with frauds and cheaters on tests and b) prevent fraud and cheating on those tests?

Liljegren: I believe that the key to preventing cheating on IQ tests lies in making the test engaging and enjoyable. People tend to cheat when they find the test boring, as they simply want to finish it as quickly as possible, similar to how one might skip through a dull movie. However, if the test is fun and feels like a rewarding journey, participants are far less likely to rush through or cheat.

When the test is designed in such a way that it encourages deep thought, curiosity, and cognitive growth, test-takers are naturally more invested in the process. This engagement reduces the temptation to take shortcuts, as participants are more interested in exploring and solving the problems presented. By making the experience fun and stimulating, you not only prevent cheating but also improve the quality of the data collected.

In essence, if the IQ test becomes an enjoyable challenge, much like a game or an intellectual journey, participants are far less likely to cheat and more likely to put forth their best effort. This approach ensures that the results reflect their true cognitive abilities, rather than rushed or dishonest attempts to finish quickly.

Jacobsen: What is an efficient means by which to ballpark the general factor loading of a high-range test?

Liljegren: To efficiently estimate the general factor loading of a high-range test, the test should incorporate a variety of question types that tap into multiple forms of intelligence and cognitive processes. This ensures the test measures a broad spectrum of abilities, including verbal, numerical, spatial, logical, creative, and abstract thinking. Using only one style of questions—such as rows of text or numbers—limits the scope of intelligence being tested, and can lead to a narrow, predictable response pattern.

Additionally, relying on the same question types repeatedly can result in a learning effect, where test-takers begin to predict the types of questions and answers. This skews the test's validity, as the test-taker's experience may be based more on familiarity with the format rather than actual

cognitive ability. Therefore, introducing a diverse range of question formats prevents this issue, ensuring that the test captures a fuller, more accurate measure of the general factor of intelligence.

Jacobsen: What is the most precise or comprehensive method to measure the general factor loading of a high-range test, a superset of tests, or a subset of such a superset?

Liljegren: The most comprehensive and precise method to measure the general factor loading of a high-range test is by employing a superset approach, which integrates a variety of subsets. The superset allows for a more holistic view of intelligence by encompassing a diverse array of cognitive abilities, such as numerical, verbal, spatial, and logical reasoning. This broad scope provides a more accurate measurement of general intelligence (g) because it evaluates a wide range of cognitive processes that overlap and interact.

By using a superset, the test becomes dynamic, capturing the interconnections between different cognitive domains. Knowledge from one subset can inform and enhance performance in another, allowing you to form a fuller, more nuanced understanding of a person's intellectual capacity. This approach not only reduces bias but also prevents the predictability of answers that can arise when a test is too narrowly focused.

Moreover, a superset allows for greater accuracy and robustness in general factor loading by avoiding the limitations of focusing on a single type of reasoning. By examining multiple subsets together, you provide a more comprehensive measure of cognitive ability, reflecting the complex interplay of various intellectual skills.

In summary, a superset ensures that you're capturing the full range of human intelligence, minimizing the biases associated with narrowly focused tests, and providing a more complete and dynamic assessment of general cognitive ability.

Jacobsen: What seem like the most appropriate places for people to start when taking your tests–taking into account their own skill sets, or others' tests for that matter?

Liljegren: My tests are designed to be accessible even to individuals with no prior exposure to IQ testing. The key idea is that as the test-taker progresses, their IQ naturally increases through the process. Each part of the test is interconnected, offering clues within the test itself to help guide them toward solving other sections. Rather than presenting isolated questions, the test is structured as a unified experience where everything fits together, fostering both growth and understanding as they move forward. This approach ensures that the process of taking the test is not only a challenge but also a journey of discovery.

The journey through the entire test is genuinely fun and rewarding. With each question solved, there's a sense of accomplishment and often laughter, which keeps you engaged and eager to continue. The satisfaction of cracking a question creates a sense of excitement, motivating the test-taker to push forward until they've solved it all. The tests are created with the intention of helping people increase their intelligence, not simply taking their money by leading them to believe they are correct when they aren't. This journey isn't just about testing; it's about expanding cognitive abilities in an enjoyable, engaging, and fulfilling way.

Jacobsen: What tests and test constructors have you considered good?

Liljegren: I believe that many tests I've encountered are designed not with the intention of fostering genuine intellectual growth, but rather to exploit the test-taker's desire for validation and to profit from their repeated attempts. These tests often provide immediate validation to make the participant feel correct, only to later disappoint them, leading to the common practice of encouraging a second (and often third) attempt to "fix" their results. This cycle is not about true intelligence testing but about encouraging further payments by exploiting a psychological pattern: the desire to prove oneself right and gain recognition from peers.

This type of testing is harmful because it focuses on validation rather than education. It relies on participants' egos, motivating them to pay again to prove they're capable, rather than helping them grow. This creates a cycle where the person is encouraged to rush through the test for validation, only to feel let down and encouraged to submit another payment for another attempt.

In contrast, my tests are designed to be engaging, fun, and intellectually rewarding, with the goal of fostering actual cognitive growth. The experience is meant to be so enjoyable and fulfilling that test-takers don't want to stop. The aim is to encourage them to fully immerse themselves in the test, where they are learning, exploring, and growing their IQ as they progress. The focus is not on tricking participants or manipulating them for financial gain but on offering a genuine opportunity to develop and discover new intellectual perspectives.

A good test should be an experience that challenges and encourages cognitive growth, one that leaves the test-taker with a sense of accomplishment and a desire to keep going. It's about helping them learn, not about creating a system where they're trapped in a cycle of disappointment and further payments.

Jacobsen: What have you learned from making these tests and their variants?

Liljegren: I spent three years on two different tests, working on them simultaneously and dedicating 2000 hours to each. When I submitted them at the same time, something very interesting happened: I scored my all-time high on one test, but my all-time low on the other. This experience highlighted just how unpredictable and subjective these tests can be.

Even with extensive preparation and effort, the outcome is not guaranteed. The tests are designed in such a way that, despite the time and focus invested, the results can vary dramatically depending on various factors—many of which are beyond your control. This unpredictability demonstrates that intelligence is not solely about raw effort or preparation; it also involves a complex interaction of factors, including problem-solving approach, adaptability, and the ability to navigate unexpected challenges.

Ultimately, this reinforces the notion that high-range tests are inherently unpredictable, and the experience of taking them can vary significantly from one instance to another, regardless of how much effort is put into preparation.

Some authors' tests feature repeated questions across multiple test versions, likely due to a combination of laziness and a desire to maximize profits. By identifying these repeated questions, I was able to deduce the correct answers through second attempts, as well as spot

consistent errors in the scoring of the tests I took. These errors included misspellings of my name, incorrect dates, and discrepancies in my raw scores. The recurrence of these mistakes suggests that many authors are sloppy in scoring, and I must take this into account when submitting my tests.

In such cases, I realized I need to factor in the author's state of mind during the scoring process. The outcome can vary depending on the author's circumstances—whether they are distracted, tired, or experiencing stress. To minimize errors, I must plan the timing of my submissions carefully, choosing moments when I anticipate the test author is most likely to score the tests accurately. For mail-based submissions, I also have to consider potential delays or disruptions, such as holidays or issues like mail theft or vandalism, that could impact the delivery or processing of my test. These external factors, which are beyond my control, require careful planning and preparation to ensure the best possible conditions for submitting my tests.

The realization that so many aspects of the process are influenced by factors out of my control has shaped my approach to testing. While it's frustrating, it also underscores the need to approach the testing process with patience, awareness, and strategic thinking to navigate these challenges effectively.

When creating my own tests, my primary goal is always to foster cognitive growth rather than to make quick profits. Too often, the rush to monetize IQ testing leads to burnout among authors, which in turn results in sloppy test scoring and a lack of care in the process. This is something I'm very conscious of, and I make it a point to thoroughly double-check and verify everything I do. Each test I score is done with the utmost care, knowing that inaccurate results can have lasting consequences on someone's life.

I understand how a poorly scored test can affect a person, particularly when they are already facing difficulties. A mistake on a test could contribute to feelings of inadequacy or frustration, or even worse, lead to a deeper sense of alienation. I am very mindful of this, and it's why I dedicate hours to ensuring the accuracy of each test I score. I want the experience of taking my tests to be constructive, encouraging, and enlightening for the test-taker, and to give them an opportunity to truly grow their intelligence.

Moreover, I believe that creating tests with integrity, where the scoring is accurate and fair, has a far-reaching positive impact on individuals. People should leave my tests feeling not only more knowledgeable but also more confident in their abilities, as opposed to feeling confused or disheartened by an inaccurate result.

Ultimately, the goal is always to provide an environment where learning is rewarding and enjoyable. This is why I am so meticulous about every detail in the process, ensuring the test is as much a tool for personal growth as it is an intellectual challenge.

Jacobsen: Thank you for the opportunity and your time, Patrick, and thank you additionally for your patience and forgiveness in my delays.

Liljegren: You're very welcome! I'm glad I could assist, and I truly appreciate your thoughtful words. If you ever need more help or have further questions in the future, don't hesitate to reach out. Best of luck with your endeavors, and I hope everything goes smoothly from here on out!

Marco Ripà and Roberto Enea, DynamIQ

2024-11-22

Marco Ripà gives some opening commentary on his involvement in the dynamic style high-range test. Roberto Enea talks about his work on DynamIQ, a dynamically generated spatial IQ test. Enea, inspired by Marco Ripà, aimed to create a test that minimizes the "training effect" and resists cheating, making it suitable for repeated use. Developing DynamIQ involved balancing design, coding, security, and fairness. The test generates questions dynamically with consistent difficulty across tests. Enea acknowledged challenges in norming tests due to low and self-selected high-IQ samples and emphasized the need for diverse populations. DynamIQ avoids linguistic bias and ensures privacy by anonymizing user data. Its credit system allows economical, flexible use over time.

Scott Douglas Jacobsen: As an opening question, what is your involvement in this dynamic style test?

Marco Ripà: Since I deleted all my data around early 2017, when I left the project, my memory of the test construction and normalization process is limited. Additionally, my memory isn't solid. I do recall following Paul Cooijmans's guidelines at the time, and if I'm not mistaken (we should verify this to be specific), I also communicated with him about the test norming to ensure everything was done correctly. The info I used to norm the test at the time also included the testee previous scores on reputable HRTs and supervised tests. Cooijmans and I talked about the

z-scores method between September 3rd and September 4th 2016.I used the z-scores instead of the rank equations since our norming sample was too small for the latter method.

I remember choosing circles, triangles, and squares to create a culture-fair test with geometric properties. This decision was based on my initial idea to develop a spatial version of the ENNDT, which I had previously developed with Gaetano Morelli (see this paper). Regarding the colour selection—yellow, green, and white/blank—we opted for these colours to avoid underestimating the performance of colour-blind individuals.

Here is a summary of the entire story (please feel free to ask Roberto for more details):

In 2011, I envisioned developing a dynamic high-range test featuring thousands of unique software-generated items. Thus, this would ensure that any collection would share the same norm. Initially, we planned to use OEIS sequences characterized by unique properties. A few years later, with Gaetano, we developed the ENNDT tests, as described in the paper above. However, this test proved extremely challenging. Only highly skilled individuals achieved positive scores, making it impractical for screening purposes—even among the gifted populations.

Our next goal was to create a spatial, culture-fair IQ test for the high range that wouldn't require participants to be exceptionally gifted. This was a more difficult challenge than the previous one. I enlisted another Italian Mensa member to help us. Roberto Enea joined the project, handling the programming and technical aspects, which I couldn't manage myself. This collaboration involved over six months of work.

After internal testing, the tool was ready. We began beta testing. Then I normalized it using the standard techniques described by Paul Cooijmans. However, that was the last time I engaged in such work. I can't recall the details. There were a lot of linear regression attempts and the corresponding R^2 values. I left the project in early 2017.

Subsequently, I deleted all documentation related to the test, the item projects, and the data on testees' performances, which only included initials, raw scores, and attempt numbers. So, I can't provide more technical information than Roberto, who has been the sole owner of the spatial dynamic test platform since 2017.

Nonetheless, I'm pleased to receive credit for the concept I envisioned over a dozen years ago—a true dynamic IQ test generator developed by software that is resilient against cheating (at least in the pre-ChatGPT era).

Anyway, I created the items following my initial idea of a mathematical method to combine those three geometrical shapes and Roberto wrote the program and created the tool to make the whole thing happen. I remember that the idea was to divide every shape in corners and/or sides and merge different shapes together.

Now I have to go.

Jacobsen: When did this interest in test construction truly come forward for you?

Roberto Enea: Hi Scott, thanks for this interview. Actually, I was mostly interested in solving tests rather than constructing them but I have found Marco's idea about implementing a system to automatically generate IQ tests very interesting and challenging.

Jacobsen: What were the realizations about the tests, at the time, and the need to develop yours?

Enea: At the time we started working on DynamIQ there was nothing similar meaning no systems that were able to generate dynamically spatial IQ tests. We were conscious of being creating something completely new.

Jacobsen: What was the origin *and* inspiration for the creation of the DynamIQ – the facts and the feelings? You have made some comments about it. Marco Ripà first mentioned this to me about 8 years ago with particular excitement about the 'ambition' behind this project.

Enea: This is more a question for Marco since he had the initial idea. The main idea was creating an IQ test that cannot be cheated and where the "training effect" has a lower impact so that you can take it several times without loosing accuracy. That is a great idea because it could be used to monitor the actual efficacy of brain training systems. The initial idea was making DynamIQ also something that could be used by professionals and academics but we never accomplished this final step.

Jacobsen: Any word of credit to others who helped in the development of this test?

Enea: Unfortunately, not. DynamIQ has been developed only by Marco and me

Jacobsen: How does design and coding play into the construction of DynamIQ?

Enea: I would say 50% designing and 50% coding. Design is not just about the test design but it includes other aspects like security, anti-cheating, data privacy, etc.

Jacobsen: What skills and considerations, in an overview, seem important for both the construction of test questions and making an effective schema for them?

Enea: The main challenge of creating a dynamic iq test is defining a sort of generation rule for test questions rather than defining the single test question, because you have to automatically generate tests whose difficulty should increase during the test (let's say from the 1st question to the 25th) but it should also be almost the same for the same question across different tests (the question n. 13 should have almost the same difficulty across all the tests generated). This requires a deep understanding about how spatial tests work "behind the scenes", meaning that you have to know what makes a spatial test more difficult than another one beyond the intuition of it.

Jacobsen: How do we know with confidence listed norms are, in fact, reasonably accurate on many of these tests?

Enea: In my opinion the answer is that we cannot be 100% sure. Most of the time the norms are computed by the authors and they are not verified by other peers. Unfortunately, until there isn't a rigorous scientific validation of the test you cannot be sure about the accuracy declared.

Jacobsen: What are the most appropriate means by which to norm and re-norm a test when, in the high-range environment so far, the sample sizes tend to be low and self-selected, so attracting

a limited supply and a tendency in a type of personality? Pragmatically speaking, for really good statistics, what is your ideal number of test-takers? You can't say, "8,126,000,000."

Enea: Rather than being the number of test takers the problem is the sample composition. For example, selecting people in the high range is not difficult. There are a lot of test takers coming from High IQ societies like Mensa who have an official assessment of their IQ. It is much more difficult selecting people in the low and middle range because in a lot of countries like Italy IQ assessment is not a common practice so most of the people are not aware of their IQ. For this reason, it is difficult to collect a sample that is homogeneous enough to actually represent the whole population.

Jacobsen: Is English-based bias a prominent problem throughout tests? Could this be limiting the global spread of possible test-takers of these tests rather than limiting them to particular language spheres?

Enea: About DynamIQ that is definitely not a limit since it is a spatial test. No language knowledge is required.

Jacobsen: How do you ensure protection of the "privacy and personal information" of test-takers? Why "share information about your use of our site with our social media, advertising and analytics partners"?

Enea: About this I would like to clarify that sharing information about website usage does not mean sharing results. The website collects anonymous information about the provenance of the visitors and other information that are usually useful for marketing but these are not related to the test itself. The test is anonymized also because we don't collect any personal information of the user. We store the email used for the registration but there is no information stored that can connect the email to the real person.

Jacobsen: What is the purpose of the account and credit system of the DynamIQ test setup?

Enea: The credit system allows you to buy several tests executions in one shot so that the user can save some money. The credit never expires so you can take the test even several years after the purchase.

Jacobsen: With the advent of the internet, cheating on individual questions and on whole tests is a possibility and a reality on these high-range tests. How do we prevent such occurrences? Also, things like the law and ethics.

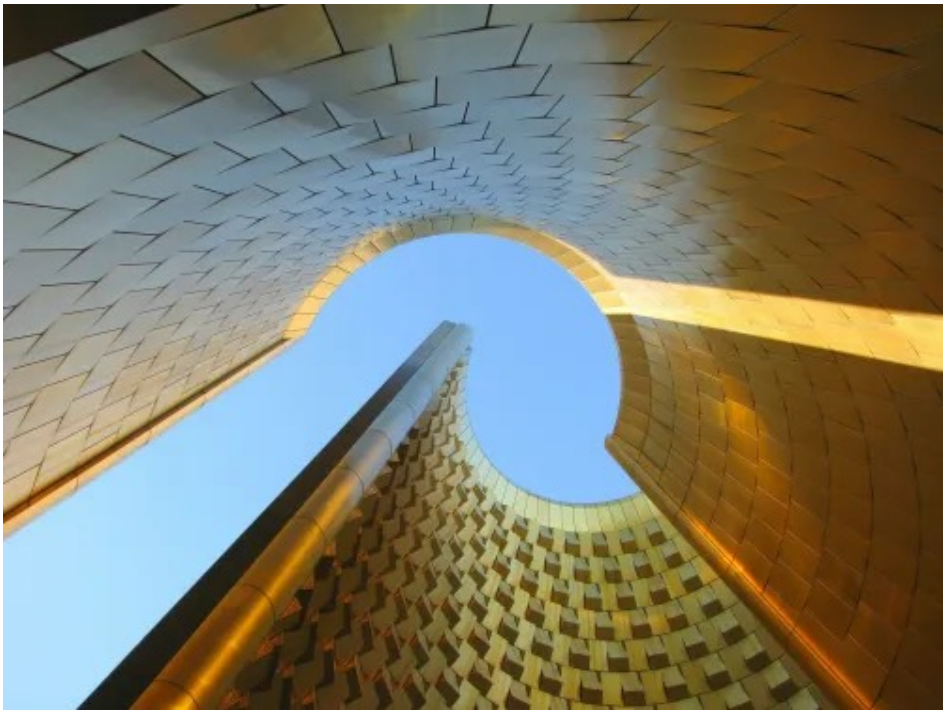
Enea: That is actually the main purpose of DynamIQ: avoiding cheating in IQ tests. The number of combinations you can have avoids that the user can somehow "memorize" the answers. Of course, in the age of AI, it might be possible to cheat using systems like ChatGPT but at least for now it seems that they are not smart enough to solve this kind of tests. There is a challenge in place called ARC prize designed by researchers at Google that is focused on creating AI model to solve spatial tests. The tests designed in the ARC prize are quite simple, not comparable to DynamIQ. Nevertheless, the best result so far is 42% accuracy meaning that the best model fails almost 60% of the times.

Jacobsen: Thank you for the opportunity and your time, Roberto.

Enea: My pleasure.

Bob Williams, The Flynn Effect: A testing phenomenon, not psychometric g

2024-11-22



Original authorship December, 2021.

The Flynn Effect (FE), characterized by consistent increases in IQ test scores over time, has been observed globally but varies significantly across nations and demographics. Initial studies highlighted these gains, with later research attributing them to environmental, behavioral, and methodological factors rather than changes in general intelligence (g). Notably, FE gains are higher in fluid intelligence measures than crystallized ones, vary by age and test type, and sometimes reverse, as seen in several developed nations. These reversals point to the saturation and decline of positive factors, coupled with the influence of negative causes such as dysgenic fertility. Analyses suggest the FE operates on non-g factors, with minimal evidence linking it to actual intelligence improvements. Methodological artifacts, including test-taking behaviors and scoring techniques, contribute significantly to the observed gains. Future research, leveraging genetic markers and polygenic scores, may further elucidate the complex interplay of factors underlying the FE's variability and reversals.

Background

The thing we now call the Flynn Effect was initially discovered by researchers in the 1940s as an increase in IQ test scores. Papers reporting such gains were published by Smith (1942), Tuddenham (1948), Lynn (1982), and Flynn (1984). This effect did not have a name until the publication of *The Bell Curve* in 1994. Herrnstein and Murray named it the Flynn Effect (see page 307). Subsequently, researchers began to look at the effect and have since published a huge

number of papers that attempt to make sense of what is happening. They found that gains were large enough to be of concern. If real, they suggest a large change in intelligence; but if not real, they at least reveal an instability in IQ tests.

Examples of IQ test score increases per decade: U.S. 3.0 points; Japan 7.7 points; and Argentina 6.9 points. Imagine a 50 year span… these gains would amount to over two over standard deviations (a very large difference). James Flynn initially noted that these gains are so large that it would mean that the average IQ in the United States in 1918 would have been 75, if scored against the norms at the time of his writing. Various similar observations (including Dutch data) showed that the gains are unlikely to be real, yet when the public and pop science magazines heard about the effect, they assured us that mankind was becoming brilliant. Clearly that was not happening, but even some researchers began to suggest that people were getting smarter.

One early question was whether the FE was real? In terms of something that can be shown to be related to another intelligence related measure, is the effect more than random differences in data? Rushton used principal components analysis to look at gains on the WISC-R and WISC-III and found a cluster, meaning that the gains were a reliable phenomenon. The cluster was independent of the cluster formed by breeding group differences, inbreeding depression and *g* loadings, which tells us that the gains are not a Jensen Effect (meaning that they are not *g* loaded). Other researchers showed a similar result by using the Method of Correlated Vectors. [01]

Today it is common for researchers to accept that the FE is a score gain of about 3 points per decade. But when we look at the changes in scores on a nation by nation basis, we find gains that are much higher and much lower. This tells us that whatever is causing the changes in one place is acting differently or is due

to a totally different cause from that in a place where the changes are quite different in magnitude. We have negative FEs (reversal) in at least seven nations–again pointing to different causes or different stages of individual causes.

The message that will emerge from this discussion is that the FE is not a single thing, but is the sum of many parts that vary over time and place. If we compute a FE in one nation it will be different in both magnitude and characteristics from a similar computation in a different nation at the same time. But if we compute it in one nation at one time and then compute it again in the same nation but at a different time, the result can be different in magnitude, sign, and component causes.

Characteristics

To get an idea of how inconsistent the FE has been let's examine how it has played out in different studies:

- Gains mostly in the low IQ range; gains mostly in the high IQ range; gains uniform for the full IQ distribution.
- The effect is seen in preschool children; some papers argue that the FE is caused by education. • Different age groups, within the same study, can show different FEs.

- Gains using the same test are higher for measures of fluid intelligence and lower for crystallized intelligence.
- Different tests give different FE changes. Many references point to the Raven's Progressive Matrices (RPM) test as showing the largest gains.
- When FE gains have been tested for invariance, the result is consistently that invariance is not supported between age cohorts. This importantly means that IQ tests operate differently for different age groups.
- Some component measures, such as spatial ability show gains, while others, such as vocabulary show losses.
- Gains seem to be in ability differentiation, not in *g*. [02]
- Some studies (Northern and Central Europe) show a significant sex difference, with larger FE for women than men.
- The FE is regional, showing larger gains in regions experiencing rapid development. • Within individual nations, some show rapid FE gains, then slow gains, then no gains, then a reversal (IQ scores declining).

Reversal

It is important to consider the cases in which large FEs declined and then reversed over a period of years. The reversals are difficult to explain by most of the causes that are otherwise plausible. FE gains have turned into losses in Norway, Denmark, Britain, Netherlands, Finland, France, Estonia. None of these nations show parallel effects that might relate to declining nutrition, physical traits (height), education, less complex environmental stimulation, etc. Woodley et al. reported a literature search that identified reported negative FEs in 13 nations. This study reported more rapid FE declines when less *g* loaded tests were used and identified immigration from low IQ nations as contributing to the net IQ decline.

How can the negative FE be explained? The answer lies in the FE consisting of numerous causes, with varying effect sizes and different saturation points. These effects reach their maximum effect and then cease to cause changes (up or down). When the causes that increase test scores decline to insignificant

levels we are left with negative causes that are still active. One known negative cause is dysgenic fertility (bright people having fewer children than dull people). This effect seems to be continuing at a slow, but steady rate in developed nations. The dysgenic effect will be discussed after the positive causes are considered (below).

Some researchers have found that the negative FE is even larger than the positive FE. Pietschnig & Gittler found a 4.8 point per decade decline in German-speaking nations. They attribute the reversal to saturation of positive FE factors. Dutton & Lynn found a 3.8 point decline in France over ten years. Platt, et al. reported a large U.S. study that showed a positive FE for IQs above 130 and a negative FE for IQs below 70 (all from the same data). In a separate study, Woodley

reported a loss of 4.5 points per decade in the Netherlands. These negative FEs are larger than the often claimed average FE gain of 3.0 points per

decade.

What causes the FE?

Various papers have investigated what they describe as THE causes of the FE. If they found some supporting evidence, they have typically presented it without noting that there are obviously many other likely contributors. Some of the things that have been considered as candidate causes:

- Education
- Decreased family size
- Increased exposure to testing
- Heterosis
- Exposure to artificial light
- More complex visual environment
- Nutrition and improved health care
- Child rearing practices
- Abstract reasoning
- Speed of test completion
- Slower life history speed
- Testing artifacts

Among other potential causes, migration, fertility, and mortality have been investigated and found to not show correlations with the FE.

Education

More years of education is supported as a cause in some studies; some researchers argue that it has the largest effect. There are, however, effects that are opposed to this cause. Numerous reports show declines in Gc (crystalized intelligence) and increases in Gf (fluid intelligence). Education should show gains during school years, but some studies have found larger gains among adults. Other studies have found that both Gc and full-scale gains were negligible, while Gf shows gains. This is opposite of what would be expected from education driven gains. As previously noted, the FE has been shown for preschool children. They have shown IQ gains of 3.9 points per decade (higher than the often stated average FE gain of 3 points per decade. This range of different findings is typical of attempts to verify specific causes.

Decrease in family size

Smaller family sizes would cause a gain in mean scores because it would disproportionately remove more people with slightly lower IQs and retain those with higher IQs due to the birth

order effect. The (related) well established negative correlation between IQ and fertility rate is the focus of study for the decline in *g* that has been studied extensively. In Iceland polygenic scores [03] were used to predict educational

achievement and showed a negative correlation with Icelandic and US data. This cause is convincingly established and points to a decline that is a Jensen Effect. [04]

Increased exposure to testing

Arthur Jensen pointed (*The g Factor*) to increased test-wiseness related to more frequent testing in schools as a non-*g* factor in increased test scores. One of the most convincing demonstrations of this came from the 72 year range of tests in Estonia. [Olev Must, Jan te Nijenhuis, Aasa Must, & A. van Vianen, (2009). Comparability of IQ scores over time. *Intelligence*, 37, 25–33.] When I discussed this with Olev Must, he

told me that one of the good outcomes of the communist period was that they never threw any documents away. Hence, they had National Intelligence Test results for this long period. Analysis of the results showed a clear trend of increased guessing (more test items tried and more errors, but also with the expected gains). This effect was predicted by Chris Brand in 1996. He wrote: "The correct strategy for testees is: When in doubt, guess." Today this testing artifact is known as the Brand Effect. Michael Woodley insightfully noted that gains that had been described as Jensen Effects, based on subtest scores showing more gains on more *g* loaded test items, could be explained as Brand Effects. The more *g* loaded subtests are also more difficult and are much more likely to involve increased guessing.

Heterosis

Mingroni argued that broadened ranges of breeding (to villages that were far enough away to be outside of the breeding group in consideration) would account for a larger gene pool that could lead to increased intelligence. Since this would be a genetic effect, it should show up (if real) as a gain in *g*. His explanation was offered with the observation that environmental effects on intelligence are small, [shown by MZ twins reared apart and adoption studies] so there must be something else happening. Of course, there is–testing artifacts, such as the Brand Effect. The heterosis explanation is consistent with secular trends in height, growth rate, myopia, asthma, autism, ADHD, and head circumference. But the effect has not been observed and it is inconsistent with FE gains in Europe before increased immigration. The developmental gains are inconsistent with IQ gains in various nations.

Exposure to artificial light

The basis of this suggestion (from Jensen) is that the pineal gland can be stimulated in animals (poultry farms do this), causing faster maturity and increased metabolism. In humans, there is little doubt that we have experienced increased amounts of artificial light from area lighting, computer screens, and television. There is, however, no data reported on this potential effect, so it cannot be accepted until a proper study shows that it is actually linked to the FE.

More complex visual environment

There is no doubt but that our environments have become more complex with the development of advanced communications, video streaming, computers, smart phones, and ever increasing automobile features. Some researchers have suggested that these environmental factors have led to changes that contribute to the FE. Armstrong and Woodley reported a significant correlation between rule-dependence and FE gains that mimic the gains seen in retesting (gains on specificity). One obvious appearance of this is in progressive matrices tests, which have been shown to be subject to learning, not only from repeat testing, but also from progressing through the test. Tests such as the Raven's Progressive Matrices (RPM) show a maximum *g* loading only when first encountered. This general effect, of learning rule based processes, exists throughout our increasingly complex environment.

Nutrition and improved pre-natal health care

It is a virtual certainty that our food and health care (specifically pre-natal) have had direct impact on birth weights, height, and developmental quotients (DQs). Richard Lynn has published several papers showing the rather rapid advances in these physical measures and has implied that they translate into IQ gains. His argument makes sense, particularly in connection with head size, which is positively correlated with skull

size and brain size. There are many studies showing the positive correlation between brain volume and IQ. When high quality IQ tests are used, this correlation is about r = +0.40. In 2018 researchers determined that the cause of this correlation is lower neurite density, that promotes more efficient neurite orientation, and more complete arborizaion in larger brains. This means that larger brains are more efficient.

The nutrition argument, as with most FE outcomes, has problems. Nutrition, as it relates to vitamins, supplements, etc. have not been shown to improve intelligence in developed nations. [In undeveloped nations insufficient intakes of iron, iodine, and folate have been found to depress intelligence.] The nations presently experiencing a negative FE have not shown nutritional decline. Gains in IQ due to these factors would make sense if they were linked to IQ gains in the lower half of the intelligence spectrum, but the gains in such things as height have been concentrated in the upper half. Flynn argued that height gains were not happening at times when IQ gains were observed.

The primary supporter of FE gains in this category was Richard Lynn. His papers discuss DQs and the other physiological factors that have been linked to improved nutrition. The strong implication from these papers is that the FE gains he has suggested are gains in *g* because *g* is known to be most strongly related to the biological aspects of intelligence. The curious thing is that Lynn has also argued that psychometric *g* is decreasing due to the dysgenic consequences of high fertility among dull people and low fertility among bright people. [*Dysgenics: Genetic deterioration in modern populations*] His arguments are vectors pointing in opposite directions.

Child rearing practices

The inherent problem with explaining FE gains or losses as the result of child rearing practices is that the FE has been found in essentially every nation that has been examined, despite large differences in child rearing practices. Additionally, adoption studies have shown that adopted

children reach adulthood with a zero correlation between their IQs and those of their adoptive parents and adoptive siblings. In short, the shared environment does not impact adult intelligence. [There is a temporary shared environmental variance that vanishes around age 12.]

Abstract reasoning

As previously noted, FE gains have been larger in tests of abstract reasoning than on tests of Gc. The RPM has consistently showed a substantial positive FE. When tests are evaluated, the item level difficulty increases as a function of the abstractness of the item. As discussed above, increased difficulty can lead to increased guessing that results in a FE gain. Another result is that when tests are compared over time, the more abstract words show lower miss rates over the time range being evaluated. This result supports Flynn's interpretation that the FE is driven (at least in part) by increased abstract thinking ability.

Speed of test completion

The Brand Effect is the result of increased guessing, but there is another related effect due to the behavioral trend of students taking the tests faster. Younger cohorts work faster. Increased test taking speed results in more test items attempted, more missed and more with correct responses. The change in speed of test taking results in a significant lack of invariance. Shiu et al. showed a 38% difference in item functioning between age groups. Must and Must showed that when invariant test items were examined, there was little or no FE. When speeded items were examined there was a large, positive FE. [Speediness is determined at the subtest level by the fraction of test items that were not attempted.]

Slower life history speed

Michael Woodley and various co-authors have argued that the FE is related to slowing life history speed. This concept is related to environmental gains in safety, food supply, and other survival needs. As living conditions improve, people are inclined to shift their priorities towards such things as education, nutrition, age when first child is born, smaller families, wellbeing, and lifestyles. This model is functionally similar

to a movement from r-strategy (more offspring and less protection of young) to K-strategy (fewer offspring and significant parental protection). When populations are maturing in favorable survival

conditions, they move from fast to slow life history speed. This shift is accompanied by lower fertility rates, more education, and improved nutrition; all of these could contribute to changes seen in the FE.

Woodley, noted that life history speed is not a genetic effect, but rather a behavioral change. [Michael Woodley (2012). A life history model of the Lynn–Flynn effect. *Personality and Individual Differences*, 53(2), 152–156.] In the context of the FE, this is consistent with various demonstrations that the FE is not a Jensen Effect. [04]

In various places and time spans, it is reasonable to claim that there are changes related to slower life history speed. This is consistent with FE gains and societal behavior. But with at least seven

European nations (all highly developed) showing a FE reversal, the life history speed model would presumably have to show a reversal (faster LHS). This reversal has not been evident.

Testing artifacts

We have already looked at the Brand Effect (increased guessing) and test taking speediness as causes of the FE. Some researchers were mislead to believe that they were seeing increases in g, when they were actually seeing different rates of guessing as a function of *g* loading and item difficulty. Another artifact, directly related to IQ tests is the use of classical test theory (CTT) instead of item response theory (IRT). Most IQ tests are scored using CTT. This method applies equal weight to each test item and simply combines subtest scores to produce an IQ score. One obvious problem with this approach is that it gives equal weight to easy and difficult test items. IRT is based on item level difficulty, as determined by the item characteristic curve. IQ can be determined by establishing the level of item difficulty beyond which guessing is indicated. IRT is understood to be the superior method.

Beaujean and Osterlind scored the National Longitudinal Survey of Youth data set using both CTT and IRT. Results are shown below:

Peabody Picture Vocabulary Test-Revised

CTT FE of 0.44 points per year

IRT FE of 0.06 points per year

Peabody Individual Achievement Test-Math

CTT FE of 0.27 points per year

IRT FE of 0.13 points per year

These results do not need explanation. They are substantial and are entirely the result of scoring the same test results using CTT and the superior IRT.

Another artifact is present in numerous studies; it is that the FE is measured at two different times, using different tests or different revisions of the same test. These differences introduce measurement errors due to different test items being used and practice effects when the same items are used. There is no literature that has sorted out the impact of this category of error, but the qualitative aspects of these are obvious and most likely relate to the inconsistent and confusing outcomes that are common in FE literature.

Are FE changes *g* loaded?

Perhaps the most important factor to be established about the FE is its *g* loading. If it is a change in *g* (a Jensen Effect), then we would have real increases in intelligence. If it is not *g* loaded, then the changes are in something else; this could be changes in non-*g* factors that relate to intelligence or simply artifacts that should be treated as noise.

Most researchers have tried to determine if the effect they were examining is a Jensen Effect or not. Almost all have found that it is not *g* loaded and it is likely that those who claimed a change in *g* were mistaking a Brand Effect for intelligence gains. As mentioned in the background

section, one of the most obvious ways to appreciate that the FE is hollow is to consider the magnitude of changes that have been reported in various nations. Over relatively short spans of time the FE gains have been outrageously large, suggesting that past generations were at the level of retardation as compared to present populations. Nothing we have seen in real world behavior is consistent with such a massive change in intelligence.

The only confirmed FE changes have been those associated with environmental effects. We already know that nothing in the environment has been found that actually increases intelligence; ergo, FE gains would not show *g* variation if they are caused by the environment. At this point in time we can safely say that the primary factors contributing to the FE are environmental (including behavioral).

An excellent study of the FE and a biological marker was done by Nettelbeck and Wilson in Australia. Two studies were done at different times (1981 and 2001). The studies were done in the same school and same grade levels using the same test (Peabody Picture Vocabulary Test). They also measured inspection time (IT) on both occasions, using the same Gerbrands tachistoscope. It is an excellent biological intelligence marker. The results showed the predicted FE gains (5 points) over the 20 year period, but the IT results were unchanged. This is exactly what would be expected if the gains were not a Jensen Effect. I asked Nettlebeck if there were any observable differences in SES or nutrition between the two groups. He said that the area served by the school was stable and that there were no observable differences in such things as nutrition or standard of living.

Principal components analysis of FE gains (discussed above) showed that there was no overlap between FE gains and purely genetic factors (racial differences and inbreeding depression). Must et al. used the method of correlated vectors [01] to test for *g* loading and found no *g* loading.

Jensen presented a particularly convincing argument that shows another way to demonstrate a lack of *g* loading in FE changes. He stated that the definitive test of whether FE gains are hollow or not is to apply the predictive bias test. This means that two points in time would be compared on the basis of an external criterion (real world measurement, such as school grades). If the gains are hollow, the later time point would show underprediction, relative to the earlier time. This assumes that the later test has not been renormed. In actual practice tests are periodically renormed so that the mean remains at 100. The result of this recentering is that the tests maintain their predictive validity, indicating that the FE gains are indeed hollow.

Finally, there has been a dysgenic effect on intelligence in developed nations for the past 100 to 150 years, caused by the negative correlation between intelligence and fertility rate. This effect is shown by measures that load on *g* (reaction time, vocabulary, color sensitivity, and backward digit span). These measures have shown movement in the direction that indicates lower intelligence. [See At Our Wits' End: Why We're Becoming Less Intelligent and What It Means for the Future, by E. A. Dutton & M. A.

Woodley of Menie. Exeter, UK: Imprint Academic.] The rate of decline in *g* is slow, but its existence means that *g* is not increasing, since this is a single parameter that cannot show a net gain and loss over the same period of time. [Also see Lynn, R. (2011). Dysgenics: Genetic

deterioration in modern populations (revised ed.). London: Ulster Institute for Social Research and Herrnstein, R. J., & Murray, C. (1994). The Bell Curve: Intelligence and Class Structure in American Life. New York: Free Press.] Besides these books there is a large number of scholarly papers also showing a decline in *g*.

Understanding non-*g* effects

IQ tests measure variances in *g*, non-*g* residuals of broad abilities, and uniqueness (specificity + random error). [Specificity = *s*, random error = *e*] The sum of these variances must equal 100%. The FE appears to operate on residuals and uniqueness. Causes such as the Brand Effect, faster test taking, and method of scoring do not change any aspect of cognitive ability, so they are confined to the uniqueness variance. That leaves a broad number of candidate causes to necessarily appear as increases in non-*g* parts of broad abilities.

If an IQ test is given to a large group, then factor analyzed (with hierarchical factor analysis), the result is that factors are shown for numerous narrow abilities; these are usually called Stratum I. When these are factored, they produce a few broad ability factors at Stratum II. The common variance in the Stratum II factors defines *g* at Stratum III. [Some tests can produce up to 4 stratums and others may produce *g* at Stratum II.] If *g* is factored out of Stratum II, the residuals are orthogonal to *g*. If these are tested for external validity, essentially none is found. The ability of IQ tests to predict important life outcomes is almost entirely the result of the test *g* variance.

In the case of FE gains and losses, the test scores are reflecting changes in the variance due to these non-*g* factors. The presence of the group factors (Stratum II) was known by Spearman and researchers since his discovery of *g*. Group factors are real abilities, even after *g* is removed. In that sense, they can and do show up in tests, causing drift up or down as environmental factors are expressed as non-*g* variance. These factors have been carefully studied with respect to score changes due to education. Learned material may show up as specificity variance, if the test calls upon such material. Another related cause of *s*-loading is test familiarity, seen when the same test is re-administered. Gains from familiarity with the test are not gains in intelligence, but can show up as *s*-loading.

Future research and polygenic scores

In The *g* Factor, Jensen discussed an idea he called an anchor point. This would be a true biological marker of intelligence (g). [The discussion (above) of IT by Nettelbeck and Wilson can be regarded as a comparison of FE gains against an anchor point.] If psychometric scores increased, they could be measured against the anchor to show that they are or are not increases in *g*. The anchor would not move if the psychometric scores were hollow. If the anchor increased, there would be a gain in real intelligence. The measurement Jensen suggested was RT (reaction time, a chronometric measure). At this point, it is fairly obvious that the FE gains are hollow, but Jensen's idea can now be done genetically by recording polygenic scores for groups being studied. If the polygenic scores increase, we have a direct measure of a change in real intelligence. Monitoring polygenic scores would also serve to confirm or disconfirm the decline in *g* that has been discussed by Dutton / Woodley and Lynn. Given the huge increase in genome

data banks, it is inevitable that such data will be used in the future to give excellent indications of real population changes in intelligence.

Conclusions

- FE gains and losses are due to an unknown number of small causes that may appear in different combinations at different times or different places.
- Gains and losses are not Jensen Effects and as such do not represent changes in real intelligence. • Reversal happens when negative causes (lowering intelligence) are larger than those causing gains. This happens when the causal effects reach saturation.
- Causes of the test score instability are associated with the environment and with test artifacts.

Notes

[01] The method of correlated vectors is used to determine wether an external variable is related to *g*. It is a somewhat complex method that is fully explained in Appendix B of Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger. The basic process is to create a column vector from the *g* loadings of subtests and then correlate that with a vector that consists of measurements of a factor that is external to the test. If there is a positive correlation, then the external variable is *g* loaded. There have been papers that challenge the merits of this method as valid for all situations. It is, however, widely used to demonstrate that various measures are related to *g*.

[02] Broad abilities (typically Stratum II factors in a Cattell-Horn-Carroll test) can be divided into *g* and non-*g* parts. In determining the *g* loading of a test, *g* is the common element in the Stratum II factors. If *g* is factored out of the Stratum II factors, the non-*g* parts can be identified as residuals of each broad ability.

These residuals are real abilities, but typically show little, if any, predictive validity when tested independently from *g*. At high levels of intelligence, Charles Spearman (who invented factor analysis and discovered *g*) contended that the differentiation between cognitive abilities shifts towards increased importance of the non-*g* (residuals). This is known as Spearman's Law of Diminishing Returns and remains in dispute because it is vexingly difficult to prove. The use of "ability differentiation" in the document is a reference to the non-*g* broad abilities.

[03] Polygenic scores – The success of genome wide association studies resulted in the initial identification of 1,271 single nucleotide polymorphisms associated with intelligence. These variants have been used to create polygenic scores, which can be used to measure IQ from the number of these that are present in the DNA of a given person. See: Using DNA to predict intelligence; Sophie von Stumm, Robert Plomin; *Intelligence* 86 (2021) 101530. Also see: Robert Plomin – Blueprint: *How DNA Makes Us Who We Are*, Penguin Books Ltd., 2018, ISBN 9780241282076.

[04] Jensen Effect – An effect that is related to *g* is considered to be a Jensen Effect. Because *g* can be used as the very definition of intelligence, a Jensen Effect means that the thing

being observed is related to real biological intelligence, and not to an artifact or factor that is not *g* loaded.

Dr. Kristóf Kovács on Accuracy in IQ, Intelligence, and Cognitive Abilities

2025-01-22

Received: January 10, 2025
Accepted: N/A
Published: January 22, 2025

Updated January 27, 2025.

Abstract

This interview includes a detailed conversation between Scott Douglas Jacobsen and Dr. Kristóf Kovács, a Senior Research Fellow and Lecturer at the Institute of Psychology and the Department of Counselling and School Psychology. Dr. Kovács leads the Cognitive Abilities Lab, focusing on research in cognitive abilities, intelligence, psychometrics, and their measurement. He critiques the limitations of IQ tests in assessing creativity, sensorimotor skills, or interpersonal abilities, emphasizing the need for detailed profiles for diagnostics over societal "IQ fetishism." Dr. Kovács explores the importance of ethical and transparent research practices and provides a nuanced understanding of IQ scores and their applications. The discussion includes the historical context of IQ testing, its practical applications, and the sociological implications of the g-factor as a statistical construct.

Keywords: Cognitive Abilities, Diagnostic Context, Educational Interventions, Fluid Reasoning, IQ Distribution, IQ Fetishization, IQ Measurement, IQ Tests, Multiple Intelligences, Percentiles, Psychometrics, Sensorimotor Abilities, Standard Deviation, Working Memory

Introduction

The document features an engaging interview with Dr. Kristóf Kovács, conducted in 2025 by Scott Douglas Jacobsen, as a recommendation from Björn Liljeqvist, former chair of Mensa International. Dr. Kovács, a Senior Research Fellow and Lecturer at the Institute of Psychology and the Department of Counselling and School Psychology, shares his insights on the measurement of intelligence, cognitive abilities, and psychometric tools. Leading the Cognitive Abilities Lab, Dr. Kovács critiques the limitations of IQ tests, emphasizing their inability to measure creativity, sensorimotor skills, and interpersonal abilities. He highlights the importance of providing detailed diagnostic profiles rather than relying on singular IQ scores. The interview delves into societal misconceptions, such as "IQ fetishism," and clarifies the statistical construct of the g-factor, noting its utility in sociological studies but limited relevance for individual diagnostics. Dr. Kovács' work underscores the need for ethical and transparent research practices and the refinement of tools to better capture the complexities of cognitive abilities. His perspectives challenge conventional views on intelligence testing and advocate for a more nuanced understanding of cognitive profiles for practical applications, ranging from education to legal contexts.

Main Text (Interview)

Interviewer: Scott Douglas Jacobsen

Interviewee: Dr. Kristóf Kovács

Section 1: Introduction and Context: Setting the Stage

Scott Douglas Jacobsen: So, today, we are here with Dr. Kristóf Kovács. This interview is a recommendation from Björn Liljeqvist, so thank you, Björn. I interviewed with him a while ago. I have been interviewing many individuals from various groups, including Mensa. In high-IQ communities, I wanted to get a professional opinion about testing. So, I posed the first big question that people might have if they are stumbling upon this interview: How much do IQ tests measure intelligence? What is the overlap between IQ and intelligence? In other words, what is the overlap in this Venn diagram?

Section 2: Defining Intelligence: Beyond the Traditional Views

Dr. Kristóf Kovács: That is a very old question. Whether IQ tests measure intelligence is a controversial issue. I do not think it is a particularly useful question because, to a large extent, it depends on how we define intelligence. If intelligence traditionally meant some form of cognitive ability, then today, with enough research, one can find references to all sorts of intelligence.

There is a paradox I perceive here. People who are very critical of IQ tests and the concept of intelligence argue that IQ testing is flawed. Yet, simultaneously, they are quick to embrace the term intelligence. There is always an alternative concept proposed to counter IQ. The first major alternative was emotional intelligence, which, after 20–25 years of research, became a meaningful scientific construct, in my opinion. However, it does not necessarily need to be called intelligence—it could be termed emotional ability. Nevertheless, now we see references to concepts like spiritual intelligence, naturalist intelligence, and other types of intelligence.

Of course, IQ tests clearly do not measure intelligence if intelligence is defined broadly enough to include aspects such as one's relationship to spirituality. IQ tests do not assess spirituality, emotionality, one's connection to nature, interpersonal skills, self-awareness, or other qualities often labelled as intelligence today. Therefore, the extent to which IQ tests measure intelligence depends entirely on how intelligence is defined. Debates over definitions, in my experience, are not particularly useful.

I try to avoid using the term "intelligence" whenever possible. Interestingly, I used to work extensively with Mensa, which is probably how you found me through Björn. However, I am primarily a researcher specializing in individual differences in cognition. My academic work at the university involves a research position.

In my research, I cannot entirely avoid using the term "intelligence," particularly in contexts related to Mensa, but I prefer to frame my research interests as focusing on cognitive abilities rather than intelligence. When we discuss cognitive abilities, there is no meaningful way to include aspects like spirituality.

Section 3: Cognitive Abilities vs. Intelligence: A Conceptual Shift

My research lab is called the Cognitive Abilities Lab—it is not called the Intelligence Lab. In my work, I consciously use the term cognitive abilities because it is plural. Intelligence, by contrast, is singular. As a researcher, discussing a range of specific abilities, such as fluid reasoning or crystallized knowledge, is far more meaningful.

Working memory or perceptual speed, and so on, are more meaningful constructs than a single general intelligence. General intelligence, in my opinion, is an index derived from various specific cognitive abilities. Still, it is not an ability in itself. For this reason, I prefer discussing cognitive abilities rather than intelligence. This approach avoids the type of definitional debates you raised. That said, I don't want to circumvent the question completely.

IQ tests do a reasonable job if we define intelligence as cognitive ability. There's a famous saying from Winston Churchill that democracy is the worst form of government, except for all the others humanity has tried. When I teach this or present at conferences, I often draw a parallel, saying that IQ tests are the worst instruments for measuring intelligence—apart from all the others psychology has ever tried.

Jacobsen: That's good. A different way to frame it is from an empirical basis. If we're examining cognitive abilities, what has emerged from research over the past century or so regarding what IQ tests measure? Also, what do the tests not measure that we know fall under cognitive abilities?

Section 4: IQ Tests and Their Purpose: Strengths and Limitations

Kovács: That's an interesting question. If we consider creativity a cognitive ability, IQ tests do not measure it. Creativity is assessed using creativity-specific tests, but it is a much harder construct to define, operationalize, and measure with psychometrically sound instruments.

Sensorimotor abilities are another relatively underexplored area in cognitive ability testing, especially in young children. In my lab, we are conducting a research project on this topic. Our

findings suggest that in preschool children, sensorimotor abilities—such as balance or other basic motor skills—are strong predictors of cognitive abilities required in school settings. Interestingly, these correlations diminish after about age seven. However, in preschoolers aged four, five, and six, sensorimotor abilities are significantly linked to skills like memory and the ability to focus, which are crucial as children begin formal education.

Sensory motor abilities and creativity are two areas that, while reasonably considered cognitive, are not measured by IQ tests. IQ tests have historically focused on educational settings and later workplace applications. The military was among the first workplaces to use intelligence tests to predict achievement or trainability. What schools and workplaces require has heavily influenced the development of these instruments.

Section 5: Standard Deviation and Interpretability of Scores

Jacobsen: People researching IQ might encounter terms like standard deviation, whether 15, 16, or other values, and lists of IQ scores—highest IQ score lists, historical figures, famous people, etc. What should people think critically about when they encounter these references? Regarding some of these popularized extraordinary IQ scores, what can we reasonably say about their accuracy? Specifically, how do high and low scores relate to rarity percentiles?

Kovács: That's a great question. There are two parts here: one about standard deviation and the other about the interpretability of the range. The most common standard deviation is the 15-point standard deviation, which was established with the Wechsler scale. This is the standard IQ distribution you'll find in textbooks. IQ is typically presented as a scale with a mean of 100 and a standard deviation of 15.

Here's how it works: your raw test score is standardized, converting it into a z-score, expressing your performance in standard deviation units. Then, we assign 15 points for every standard deviation. For example, if you score exactly one standard deviation above the mean, your IQ score will be 115. If you score two standard deviations above the mean, your IQ score will be 130.

You're right, though, that other standard deviations are in use. For instance, some tests historically used a 16-point standard deviation. However, I'm unsure if that is still true with the Stanford-Binet scales. The Cattell scale, on the other hand, used to have a standard deviation of 24. As someone who has provided feedback on IQ tests, I find this variability somewhat frustrating.

Many people, understandably, don't realize that IQ is simply a relative scale. Without a background in statistics, interpreting it can be confusing. IQ is not an absolute measure.

For example, you can express even something like height on an IQ scale. You do not need to, since height has an absolute zero, so we use absolute measures like centimetres. IQ, by contrast, lacks an absolute zero—it's purely comparative. Everyone is compared to the mean, and differences are expressed in standard deviation units before being translated into IQ scores. But if you really want you can express height using an IQ-style scale. In this case it becomes a relative score. For instance, let us assume that the average height for Canadian males is 175 centimetres,

with a standard deviation of 6 centimetres. If someone is one standard deviation above the mean, their "height IQ" would be 115. This approach standardizes the data for easier comparison.

Jacobsen: Centimeters work—we're Canadian and use metric and imperial measurements.

Kovács: Perfect. So, if we continue with that example, a two-standard-deviation height above the mean—187 centimetres—would correspond to a "height IQ" of 130. Of course, this is just an analogy to explain how IQ operates as a comparative scale rather than an absolute measure.

IQ scores can always be translated back to standard or z-score scores. For example, if you're just above one standard deviation above the mean, your z-score would be +1. If you're exactly as tall as the average Canadian male, your height in a standard z-score would be 0. If you're one standard deviation above the mean, your z-score is +1. Theoretically, you could translate that into an IQ scale, but why would you? There's an absolute zero with height, so you don't need to use a relative scale like IQ.

IQ, conversely, is purely a relative scale. If you know someone has an IQ of 150 but don't know the standard deviation being used; you can't determine if it's three standard deviations above the mean or slightly less than two. For example, with a standard deviation of 24, an IQ of 150 represents something different with a standard deviation of 15. People often don't realize the importance of standard deviation in interpreting IQ scores.

Section 6: Percentiles vs. IQ Scores: Simplifying the Complexity

At the same time, there's this strange IQ fetish in society. For example, you often hear claims from celebrities—actors or actresses—saying they have an IQ of 180. These numbers are thrown around, but they lack context. In my experience, percentiles are far more useful and comprehensible for the general public.

If you have a normal distribution of scores, any z-score can be converted into a percentile or an IQ score. Theoretically, These measures are interchangeable, but percentiles are much easier for most people to understand. For instance, if you tell a parent their 12-year-old outperforms 95 out of 100 children of the same age, they will understand what that means. Similarly, if you say, "Your child has a better vocabulary than 98 out of 100 children their age," it's immediately relatable.

If you tell the parent that the 98th percentile corresponds to a z-score of +2 or an IQ of 130, it becomes more abstract. If you say their child has an IQ of 130, most people won't know how to react. Should they be ecstatic? Perhaps they read in the paper that morning about a celebrity claiming an IQ of 190, and they might feel disappointed. In reality, an IQ of 130 is excellent—it's in the top 2% and qualifies for Mensa membership.

If I were in charge, I'd eliminate IQ scores entirely and only use percentiles. In my experience, IQ scores create more confusion than clarity. Unless someone in this field understands the statistical nuances, they often misinterpret the scores. Since IQ scores can always be converted to percentiles, the latter is more intuitive and effective for communication.

On the other hand, it couldn't be clearer to a parent if you say, "Your child outperforms 90 out of 100 peers," or, "Your child is weaker than 80 out of 100 peers." That immediately highlights whether a specific area is a strength or a weakness for the child.

Section 7: Diagnostic Contexts: The Importance of Comprehensive Testing

The other question was about the range of interpretable scores. Typically, all scores are normed against a sample, usually a few thousand people. For example, in a representative sample in the U.S., you might have 5,000 or 6,000 participants, with around 200 individuals for a specific age group, such as 12-year-olds. When you compare an individual to that age group, anything beyond one in 200 is based on extrapolation.

The more you project beyond your data, the less accurate the interpretation becomes. For instance, if someone claims a child is "smarter than one in a million," but the comparison is based on only 200 children, that projection is highly speculative. Typically, scores within plus or minus two standard deviations from the mean are interpretable. A third standard deviation can also be meaningful, especially for individually administered tests that take significant time to complete.

IQ scores are often calculated as scores derived from multiple subtests. If someone scores in the top 2% across five subtests, the likelihood of that occurring across all subtests is much rarer than 2%. To explain this with an analogy: imagine you're looking for people who are taller than 98% of Canadians and have driven more miles than 98% of Canadians. The probability of finding someone who satisfies both criteria is much smaller than 2%.

Similarly, if someone scores very highly on multiple subtests, it provides a stronger basis for interpreting their overall IQ as being exceptional. By contrast, if someone scores high on just one test, that result is more likely to be "noisy," with a larger margin of error.

In statistical textbooks, normal distributions are usually illustrated up to plus or minus three standard deviations because this range covers 99.7% of the entire distribution. Only 0.3% of scores fall outside this range—0.15% on each end. For example, anything above three standard deviations would represent about 3 individuals out of every 2,000. That's why illustrations of normal distributions in textbooks typically stop at three standard deviations; beyond that, the probabilities become increasingly rare and harder to measure accurately.

Up to plus or minus three standard deviations is meaningful and reliable. I know there are groups like the higher sigma societies, but I don't want to comment. I'll leave that to someone you might interview from those societies. For the record, what I'm describing here is what you'll find in standard statistical textbooks. Reliable and valid testing generally falls within plus or minus three standard deviations. Beyond that, scores become far less reliable.

I'd be skeptical of scores above +3 standard deviations and specially above +4. A score of +4 can be equivalent to one in a million. For instance, someone claiming, "My child is smarter than 999,999 other children," raises the obvious question: how do you know?

Section 8: Multiple Intelligences and Alternative Theories

Jacobsen: These issues often tie into statistical limitations, such as sample size and whether the test was properly proctored. Then, there are potential conflicts of interest. For example, if someone takes a test designed by someone they know, the results could be biased. Setting aside those issues, we've covered a lot so far: definitions of intelligence, the scope of IQ tests, reframing to cognitive abilities, standard deviations, and reliable ranges. What about the context in which these tests are proctored? For example, tests developed with significant investment and large sample sizes are conducted in secure environments where answers aren't leaked—what is the importance of those measures when trying to measure what IQ tests aim to assess?

Kovács: In short, high stakes. Suppose you want an elaborate and thorough measurement, especially when the stakes are high. In that case, ensuring the test is secure, properly administered, and statistically sound is essential. This is particularly critical in diagnostic contexts.

One high-stakes example is the death penalty in the U.S. Individuals with an IQ below 70 cannot be sentenced to death. Determining whether someone's IQ is below this threshold becomes a matter of life and death—the highest stakes imaginable. While that's not my area of research, it's an extreme case where the reliability of IQ testing carries enormous weight.

More commonly, professionally proctored IQ tests are administered for diagnostic purposes, particularly in school settings. In the U.S. alone, millions of individually administered IQ tests are conducted yearly. These tests help identify cognitive strengths and weaknesses to guide educational and developmental interventions.

Section 9: The g-Factor: Index, Not Ability

A comprehensive profile, derived from a range of subtests, is so important. It provides a detailed view of strengths and weaknesses. For example, one of the most common recommendations by school psychologists is to suggest that a child be given extra time on tasks or exams.

Imagine a child with a profile showing excellent fluid reasoning (nonverbal problem-solving), strong verbal ability, and strong spatial ability but only slightly above average working memory and average perceptual speed. This profile often leads to frustration because the child's abilities outpace their processing speed. In other words, their strengths cannot fully compensate for the slower speed at which they process information. This kind of detailed profile allows a school psychologist to make targeted recommendations to address the child's specific challenges.

Individually administered tests are resource-intensive, typically taking one to one-and-a-half hours of a psychologist's time in a one-on-one setting. This level of investment is far greater than administering a group test to 30 students, so it's generally reserved for high-stakes situations. For instance, if a child is underachieving, frustrated, or showing signs of learning difficulties, then creating a full-ability profile is worth the investment. A detailed profile highlights individual strengths and weaknesses. It is far more useful for diagnostic purposes than a single overall score.

When I teach this, I often use an analogy to explain the limitations of an overall IQ score. Imagine visiting your doctor and receiving a detailed lab analysis of your blood sample. You see values for glucose levels, cholesterol, vitamin levels, and so on. Imagine the doctor told you,

"Your health IQ is 70." What would you learn from that? You'd know you're in trouble—only 2% of people your age are less healthy than you—but it wouldn't help you or your doctor determine what's wrong or how to address it.

That's the issue with relying solely on an overall IQ score. It's like receiving a "health IQ" score that says you're less healthy than 95% of your peers. While that might motivate you to worry, it doesn't provide actionable insights. Similarly, while overall IQ scores can be useful to an extent —such as for Mensa membership, where the goal is to identify the top 2% of cognitive performers—they don't provide the diagnostic depth necessary to understand and address specific challenges.

A health quotient (HQ) might be useful if your goal is to create a society comprising the healthiest 2% of people. However, if someone is unhealthy, an HQ score won't help them. What they need is a detailed diagnostic to identify the specific problem. That's why we use detailed tests and invest significant resources and time to assess a child individually and create a profile of their strengths and weaknesses.

Jacobsen: These are important cautionary tales about interpreting results. What about multiple intelligences, Sternberg's triarchic theory of intelligence, and the g-factor? While there's no general consensus, what is the prevailing view?

Kovács: These are all controversial topics. Regarding multiple intelligences, I think Howard Gardner's work critiques the educational system more than a true theory of individual differences. Gardner has never shown much interest in rigorously measuring these intelligences. Essentially, his theory advocates focusing on children who might not be conventionally "smart" but excel in areas like social skills or the arts. It's an example of extending the concept of intelligence, which is valuable in its own way. However, Gardner hasn't developed reliable assessment tools for most of this proposed intelligence.

Whether we should call someone "intelligent" for having exceptional interpersonal skills despite not being conventionally smart is a matter of perspective. I'll leave that judgment to others. As for the g-factor, that's closer to my area of research. My work focuses extensively on interpretations of the g-factor, and I've published on this topic. We have a framework called the Process Overlap Theory, which explains the g-factor without requiring the assumption of a general intelligence or overarching ability. Naturally, I'm biased because this is my research field. Still, I see the g-factor as a summary or index score of separate cognitive abilities.

The g-factor is statistically advantageous in many ways. While it doesn't represent a single ability, it's a latent construct useful for certain purposes. For example, suppose you're conducting large-scale sociological research and want to study how cognitive functioning predicts income. In that case, the g-factor is a highly effective tool. In that context, it doesn't matter whether someone excels in working memory, perceptual speed, or vocabulary—the overall level of cognitive functioning matters.

However, the utility of the g-factor depends entirely on your purpose. For diagnostics, the g-factor is not particularly helpful. Like the HQ analogy—it provides an overall score but doesn't tell you much about specific strengths or weaknesses. If your goal is to diagnose and support

individuals, identifying patterns of cognitive strengths and weaknesses is far more informative. On the other hand, if you're studying broad trends, such as the relationship between cognitive functioning and socioeconomic outcomes, the g-factor is invaluable.

If you want to predict someone's salary based on their cognitive abilities, overall scores or indicators like the g-factor are very useful. However, I don't see the g-factor as a proxy for a single "general intelligence." Instead, it's an index score calculated from various distinct abilities.

Jacobsen: That's a very interesting perspective. I hadn't heard it framed as an index at a sociological level rather than as a generalized commentary on a larger sociological construct. Viewing it as an index aligns with your emphasis on cognitive abilities about different factors. That makes the research clearer, too.

Kovács: Exactly. I'm glad it makes sense.

Section 10: Final Reflections: Caution and Clarity in Assessment

Jacobsen: Any final important things people should remember when they look at scores or assessments?

Kovács: That topic would take over a minute to address, so I'll leave it at that for now. If that's okay with you, my part is complete. I look forward to seeing the transcript.

Jacobsen: Excellent.

Kovács: Thank you for your time and patience.

Jacobsen: I truly appreciate this conversation.

Kovács: Thank you so much. Cheers!

License & Copyright

Author Biography

Scott Douglas Jacobsen is a Canadian author, interviewer, and publisher, and a board member and executive on numerous boards whose contributions to secularism, humanism, and human-rights discourse are distinguished by their rigour and accessibility. He established In-Sight Publishing in 2014 to produce freely available or low-cost e-books and periodicals under a Creative Commons license, thereby ensuring broad dissemination while safeguarding intellectual property.

As editor-in-chief of *In-Sight: Interviews* (ISSN 2369-6885), launched in 2012, Jacobsen curates and presents meticulously prepared, long-form dialogues with a wide range of interlocutors. These interviews include scientists and philosophers, activists and public intellectuals, addressing themes such as secular ethics, freedom of expression, evidence-based policymaking, and the global defence of human rights. His work appears regularly in peer-recognized outlets, including *The Good Men Project*, *International Policy Digest* (ISSN: 2332–9416), *The Humanist* (Print: ISSN 0018-7399; Online: ISSN 2163-3576), Basic Income Earth Network (UK Registered Charity 1177066), *A Further Inquiry*, Canadian Humanist Publications (CA Registered Charity 118833284 RR 0001), *Uncommon Ground Media* (UK Registration 11836548), The New Enlightenment Project, *News Intervention*, *Canadian Atheist*, Trusted Clothes (CN: 9562184; BN: 791402928RC0001), among dozens of others.

Jacobsen engages globally and interdisciplinarily with issues of social justice, belief plurality, and economic equity. Jacobsen has held the Tobis Fellowship in Research at the University of California, Irvine, on multiple occasions, contributing to empirical and normative studies on ethics and public discourse. He maintains active membership in numerous professional media organizations, fostering adherence to editorial standards and facilitating ongoing intellectual exchange.

His editorial leadership and commitment to open-access formats have generated a substantial, publicly accessible archive—known as the Jacobsen Bank—that documents contemporary secular and humanist thought with over 10,000 . Based in British Columbia, he continues to expand the reach of his platforms, amplifying diverse perspectives and promoting evidence-based dialogue across cultural and disciplinary boundaries.