Psychometric Intelligence Research: A Case of Degenerate Bootstrapping[1]

Gökhan Akbay[2]

gakbay83@gmail.com

Abstract

In this paper I propose a concept to describe the circular developmental trajectory of psychometrics of intelligence in the twentieth century, and I argue that this circularity explains the degenerate character of the field. Defining, measuring, and explaining intelligence formed a closed circuit of reciprocal refinement activities. I call this circular, internally guided, and non-progressive refinement process degenerate bootstrapping. Bootstrapping, especially in the initial stages of a science, is inevitable and might end up with better measuring instruments and a better theoretical foundation. In the psychometric intelligence case, the absence of truly test-independent benchmarks, over-reliance on test score correlations, and the absence of genuine theorizing prevented the field from making significant conceptual progress. The circularity is specific to psychometric intelligence research and the diagnosis of degenerate bootstrapping does not apply to neighboring fields and approaches. To describe the bootstrapping process, I will offer a conceptual history, starting with Alfred Binet and focusing on the work of American founders, namely, Lewis M. Terman and David Wechsler. Methodological and conceptual aspects of the circular modifications will be discussed with special emphasis on the definition and measurement of intelligence as well as the status of factor-analytic theories. A current brain based theory of intelligence, Parieto-Frontal Integration Theory (P-FIT) can escape this circularity to the degree that it goes beyond the conceptual confines of psychometrics.

Keywords: intelligence, psychometrics, scientific progress, validity, robustness, P-FIT

---

[1]**Forthcoming in *European Journal for Philosophy of Science***

[2]**Post-doctoral researcher, The University of Texas at Austin, Department of Philosophy, Austin, TX USA.**

Introduction

Many generations of people who grew up and lived part of their lives in the twentieth century have been subjected to IQ tests in one way or another. The idea of intelligence as a unitary, objectively measurable property of the individual had many merits in the eyes of educational reformers as well as industrialists of the early 20th century. Educational and occupational placement, diagnosis of mental deficiency were some popular uses of the tests. Intelligence testing has lost its popularity generally and within psychology in the 21st century. The negative outlook on IQ testing has influenced psychology students as well as the public. One example is Rutgers, where Louis Matzel (2024) conducted a study on psychology students, found that their views on the entire field of psychometric intelligence research were mostly negative, and they seem to be influenced by the views of the critics of "the hereditarian theory of IQ", the best known example being Gould's (1996) classic title *The Mismeasure of Man*.[3] Anti-testing movement led more than 60.000 children and their parents to refuse to take federally mandated tests in New York alone, and many other states followed course (Strauss, 2015). Majority of colleges in the US adopted test-optional admission policies, due to the raising awareness of racial and socioeconomic diversity issues (Felegi, 2024). SAT scores, which are highly correlated with IQ, are no longer required to enter many prestigious universities. Test-optional policies became especially popular as a response to Covid-19 pandemic, but many top institutions continued the practice after this period (Lovell &

[3]**One earlier and very interesting example of such a critical take on intelligence can be found in the fifth edition of the manual for WAIS (Matarazzo, 1972). The book presents the most popular IQ test, from one of the founders of the field, but handles theoretical and political problems so profoundly that it is one of its kind. It presents not just the mainstream psychometric perspective, but it also incorporates the political sensitivities of the 1960s (e.g. civil rights movement), sociological studies, brain research and twin studies related to intelligence.**

Mallinson, 2024). Introductory psychology textbooks are highly critical of intelligence testing (Warne et al, 2018). Currently, IQ tests are employed for the diagnosis of specific learning disabilities and for identifying "gifted" children, but not for rank ordering individuals in the normal range. Some companies use job-specific aptitude tests, but a test of general intelligence is not a requirement in the hiring process in the US.

One major reason for the loss of public interest was the adverse social effects of mass testing, specifically in the US. From the Johnson-Reed Immigration Act of 1924, to enforced sterilization laws, the scientific opposition to Head-Start (i.e. Arthur Jensen's 1969 paper), and more recently from the publication of the anti-egalitarian manifesto of Herrnstein and Murray (1994) and to the Clinton era welfare reform of 1996, IQ tests have been abused for a racist and classist political agenda in the US. These episodes have created a heated political and academic debate on the relations between intelligence, heredity and race (Lewontin et al, 1984; Fischer et al., 2006; Jacoby & Glauberman, 1995; Staub, 2018; Tucker, 2024). The scientific status of IQ testing has also been questioned by many researchers (Howe, 1997; Joseph & Richardson, 2024.; Murdoch, 2007; Richardson, 2000, 2022). Critiques pointed out to the lack of genuine theorizing in psychometrics, problems with measurement and definition of intelligence, cultural and racial biases in standardized tests, and the methodological problems with genetic studies of intelligence.

The main goal of this paper is to identify what has gone wrong in the scientific side of this field and to develop an abstract characterization of its evolution. Some form of stagnation is apparent. The most popular intelligence scales (SB and Wechsler) did not change much in the last century. Technical improvements in standardization or norming aside, the contents and the format are almost the same. Given that we should have learned more about the nature of intelligence, how could the tests that "define" intelligence stayed the same, asks Castles (2012, p. 108). IQ tests of today are better with respect to certain psychometric criteria, such

as being less biased and more reliable, but psychometrics does not provide a better understanding of the phenomenon it "measures". In a sense, there was technological improvement without genuine scientific progress. Can we find a general logic to this stagnation, within the philosophy of science?

One key inspiration of this paper comes from Ken Richardson (2000) who compared test construction to "a reformatting exercise: ranks in one format (teacher's estimates) are converted into ranks in another format (test scores)" (p. 31). I believe that this reformatting idea could be applied more broadly, and not just to test construction. Definition, measurement, validation, and theory construction ran on a similar, circular path which I call degenerate bootstrapping.

In the history of computer science, bootstrapping at first meant a system's coordinating its actions from within. Then, with the advent of cybernetics, it started to be identified with coupled negative and positive feedback mechanisms that govern the evolution of a system (Bardini, 2000, pp. 24-5). In this paper, I use the concept of bootstrapping as referring to a process of autocorrection or auto-improvement in a research tradition, especially at the beginning phases where the tradition builds its foundations from scratch. It is a process of self-definition as well as self-refinement. Any research tradition can begin with this type of process, with less-than-ideal definitions and less than ideal methods, improving them on the go and creating a theoretical foundation in the later stages of its evolution.

There are progressive versions of this bootstrapping process, such as the invention and improvement of thermometers (Chang, 2004), the process of the discovery of computer mouse by Engelbart (Bardini, 2000), science-technology symbiosis leading to groundbreaking discoveries in the Laboratory of Molecular Biology (Gebel et al., 2024), the discovery and medical application of penicillin (Cartwright et al., 2022). Progressive

bootstrapping is in line with Chang's (2004) "epistemic iteration" and Cartwright et al's (2022) "virtuous tangle"[4] concepts. The common properties of these processes are that multiple, less-than-ideal components (e.g. theories, measurement tools, hypotheses, organizational practices, etc.) support each other and processes of mutual correction and improvements, without an absolute foundation or preset benchmark, lead to overall progress. The foundation is built on the go, and the progressive traditions usually expand their reach by either conquering or integrating with other traditions. By these possible developments, the field would have escaped from the circularity.

Psychometric intelligence research did not progress in this manner because it could not anchor itself into stronger research traditions, and it could not build such an anchor by itself. That anchor would be a theory that connects dots and pieces of data to definitions and measurement, which also explains anomalies in a coherent fashion. For instance, in the case of thermometry, the initial measurement tool, the thermoscope was an imperfect tool because it measured a mixture of temperature and pressure (Block and Dworkin, 1974, p. 346). The ideal conditions under which a thermometer measures temperature were identified in a progressive manner where each improvement in the instrument was based on a better understanding of the phenomenon and vice versa. The long story short, measurement and theory construction coevolved in a fruitful manner in the case of thermometers, but the process was degenerate in the IQ case.

The "degenerate" of "degenerate bootstrapping" originates from Lakatos' (1978) concept of degenerating research programmes, which keep up with novel discoveries rather than

[4] **The virtuous tangle concept (Cartwright et al, 2022) has been proposed to explain the reliability of scientific products rather than progress *per se*, however, the concept has rich implications for distinguishing good scientific practices from suboptimal practices in evaluating progress-related issues. For instance, two properties of a virtues tangle, being rich and long tailed, seems to be lacking in the IQ case, but entanglement is present. Maybe this is the reason the field seems to run on a circle (i.e. entangled), and this is why it begins to progress when it relates to different fields (i.e. becomes rich and long-tailed.).**

predicting or producing them beforehand. Lakatos (1978), although not writing much on the scientific character of psychology, was considering social psychology as an *immature science* rather than a degenerating one, because only a theoretical foundation and not statistical sophistication would make a field a genuine science in his methodology. In the absence of this foundation, the condition of continuous growth cannot be satisfied. His assessment of quantitative social science in general was extremely negative:

> After reading Meehl [1967] and Lykken [1968] one wonders whether the function of statistical techniques in the social sciences is not primarily to provide phoney corroborations and thereby a semblance of 'scientific progress' where, in fact, there is nothing but an increase in pseudo-intellectual garbage (Lakatos, 1970, p. 176).

Lakatos held degenerating research programmes with much higher regard than such prescientific, data driven research practices. Degenerate research programmes are programmes with a continuity and theoretical integrity.

In Lakatos' theory, whether a programme is degenerating or not can only be determined retrospectively. Moreover, in order to make a progressive-degenerate distinction, there must be alternative research programmes in the field. It is difficult to say that there was such competition in the field of psychometric intelligence in the 20th century, except for the competing factor-analytic theories (e.g. Thurstone's primary mental abilities or Spearman's monarchic theory with *g* on top).

For all these reasons, I would like to emphasize that the term degenerate as used in this paper is not synonymous with Lakatos' concept.[5] As an adjective describing the process of bootstrapping, degenerate is used in this paper to describe a collective scientific activity that shows continuity in terms of methodological assumptions and technological development

[5] **Branahl (2024) states that in such a case with no competition, it would be better to call the programme stagnant rather than degenerating. I prefer the latter but confine the meaning of degeneracy to the consequences of bootstrapping in this paper.**

with a strong internal focus[6], that does not lead to new and surprising discoveries, and fails to penetrate the depths of the intelligence phenomena, despite technical advances in test construction and interpretation.

One important aspect of Lakatosian degeneracy is that it is not the end of the road for a research programme. Such a research programme has the potential to become progressive, once the positive heuristics are reworked to produce novelty and/or the programme reconsiders the framework by which it formulates the problems to be solved. The second option refers to problem-shifts and these can be either progressive or degenerating. According to Lakatos, a research programme might transform its original problem – e.g. explaining intelligence vs. measuring intelligence – and this shift is a degenerating one if it ends up "with solving (or trying to solve) no other problems but those which one has oneself created while trying to solve the original problem" (Lakatos, 1968, p. 317). Psycometric intelligence research has transformed the broader problem of the nature of intelligence into a problem of how to rank order individuals by a supposedly objective measurement instrument.

What could have saved psychometrics of intelligence from degenerate bootstrapping would be to tackle the connections of tests with observations in other fields such as comparative anthropology or experimental studies of cognitive functioning, to devise tests or other measurement tools based on cognitive processes, to investigate possible interventions that can make a difference in problem solving or learning abilities. Since intelligence has been considered as a biological phenomenon by the mainstream intelligence researchers, instead of ranking people, the nature of universal human cognitive abilities and their relation to cognition in other living beings could have been addressed. Individual differences could have been situated on this background. In other words, an integrative approach that favors *investigation* over *measurement and ranking* would be more fertile. Unfortunately,

[6] **I would like to thank an anonymous referee for pointing out to this "internal focus" notion.**

psychometric intelligence research was characterized by a narrow focus on tests and their statistical interpretation.

Here I examine the internal processes that led to degeneration. However, for a more complete explanation of this diagnosis, certain external factors need to be mentioned. The main external factor was the social pressure on psychology for practical utility at a premature stage of its development. Intelligence tests were invented in France as a practical tool to identify students with learning difficulties and to place them in special classes. Later, especially in the USA, mass testing was utilized in various contexts, ranging from officer selection in the army to institutionalization of the "feebleminded". The founders of the mainstream intelligence research in the US had ideological motives like technocratic progressivism, hereditarianism, racism and they had practical motives emerging from their ties to the Army, educational institutions and test publishing companies (Brown, 1992). This social relevance of testing increased the prestige of psychology in general and provided extra motivation for those working in psychometric testing. For example, Lewis Terman, as the president of the APA, defended the value of psychometric tests against experimental methods,by appealing to how they "brought psychology down from the clouds and made it useful to men" (Terman, 1924, pp. 105-106). The absence of a theory of cognitive processes underlying intelligence was discussed by key figures such as McNemar (1964), but the field continued as it was, because of the utility of tests in practical contexts. The argument of practical utility remained strong till the end of the 20th century (Jensen, 1998, p. 109). The uncertainties surrounding psychological "measurement" were known, but were excused due to this factor. For instance, Herrnstein (1973) had warned those who compared psychometrics to physics:

> "More common by far, especially in the inchoate science of psychology, whose virtue is a *pragmatic predictiveness* over some more or less limited range of events, beating out vying theories by a hairsbreadth of accuracy or elegance here and there" (p. 98, emphasis added).

“Pragmatic predictiveness” became the core virtue of the tests and their main source of external validity, which was also the basis of Herrnstein and Murray’s (1994) hereditarian manifesto, *The Bell Curve*.

Although the technical issues of test construction and interpretation and the aforementioned sociopolitical issues are intertwined in the history of psychometrics, I prefer to distinguish between sociopolitical issues and intra-scientific processes because I believe that the latter are worth analyzing in their own right and that the bootstrapping idea, if developed thoroughly, would provide a general internalist model applicable to similar cases. For this study, it will be sufficient to show that psychometrics' claims of internal coherence and external validity are the product of a circular process, so that the robust structure that emerges is in fact in appearance. To explicate the latter, I will try to show how the bootstrapping mechanism worked at critical junctures in the history of IQ tests.

The paper begins with Alfred Binet’s vacillating position on mental measurement. There I will try to show that the circularity in test construction and validation, the connection between meritocracy and measured intelligence, the tensions between detailed clinical observation and standardized mass testing begins in Binet’s work, although he did not share the hereditarian-fatalistic views of his American followers. The second section will cover the development and standardization of the tests in the US, and it will exemplify degenerate bootstrapping in test construction, creation of the so-called bell curve, and early attempts at defining intelligence. The third section will extend towards methodological problems in psychometric intelligence research, such as measurement, validity, and factor-analytic theories. I will discuss how a “pseudo-robust” justification and refinement scheme is underlying the degenerating aspects of the field. In the last section, I will discuss how one current brain-based theory – Parieto-Frontal Integration Theory (P-FIT) – inherits its

theoretical foundation from classical psychometrics but is still potentially progressive due to the richness of the observations it incorporates into its theoretical structure.

1. Alfred Binet: The Origins of Bootstrapping

By the 1890s, French psychology was gradually breaking away from philosophy and becoming an experimental and quantitative science. Alfred Binet, the developer of the first IQ test, was one of the pioneers of this school, and in the 1890s he and Victor Henri wrote articles in which they argued that instead of speculating about an abstract, unmeasurable, unobservable universal human mind, it was necessary to observe the cognitive abilities of individual human beings in all their richness (Carson, 2007, p.135; Nicolas et al., 2014). Binet and Henri preferred to focus on higher cognitive abilities, rather than simple perceptual tasks tackled in psychophysics, and preferred observing *individual differences* in performance, rather than the universal trait. The abilities they studied ranged from clearly cognitive ones such as attention, memory and understanding, to personality traits such as suggestibility. One obvious venue for practical utility of their approach was education.

Binet, had joined the committee for developing a method to distinguish intellectually disabled children from the "normals" in the schools of Paris, so that they can be sent to special education classes. His official task was to diagnose "subnormal" children, not defining intelligence, let alone ranking people. The main problems he had to resolve were to clarify the meanings of diagnostic terms concerning mental defect and to arrive at precise "descriptions of the symptoms which reveal, or which constitute a certain particular malady..." (Binet & Simon, 1905a/1916, p. 13). Physicians were using terms such as moron, imbecile, or idiot, but the meanings of these terms differed from one country to the other and even from one physician to the next. The problem was deeper than a terminological one: Binet & Simon were trying to find an objective, operational measure of the symptoms, which

would lead to “a precise basis for differential diagnosis” (Binet & Simon, 1905a/1916, p. 14). According to their classification, mental defects essentially “consist in the weakness of intelligence” (p. 22).

They defined intelligence as the faculty of judgment, which they believed to be a natural endowment of the individual, and tried to distinguish this from learned skills. They intentionally ignored the question of etiology – whether the differences reflect innate endowment or environmental exposure – to focus on what the tests say about the child’s current level of intelligence. The scale they developed was a ranking and classifying instrument. Classification of distinct levels of mental deficiency was the main goal. Although their task at first was to diagnose distinct levels of mental defectiveness, they understood that the scale could be used to rank normal individuals as well. Later in the same year, they proposed weakness in “abstract thinking” as the litmus test of subnormal intelligence, and especially for the highest grade of defectives, *morons* (Binet & Simon, 1905b, p.180).

Binet & Simon’s work was atheoretical, as Binet knew well that earlier attempts at theorizing intelligence (and mental defects) had been unsuccessful due to their reliance on philosophical speculation rather than precise measurement. Binet could not have hoped to begin with a ready-made definition of intelligence, because there was none that would satisfy his desiderata, such as measurability, diagnostic merit, and conformity with his intuitions. The crucial ideas were that the measurement of intelligence should consider the children's age and the scores should be in line with some external criterion such as teacher assessments. Thus, tests were constructed for each age group, such that the average score would be the norm for that age.

Binet & Simon continued to refine the tests, by empirical observations on children, till 1911, the year of Binet’s death. In the latest revision, they removed certain subtests because they

were either redundant or based on learned information rather than intelligence. The logic of the refinements resembles a bootstrapping procedure: begin with an imperfect but sufficiently variable set of items to probe into various aspects of intelligence, see the discrepancies in results (e.g., scores not conforming to an external criterion like teacher assessment), make the corrections on the tests and administer them again. They also started to use the scale as a *ranking instrument and not just a diagnostic one.* They even spelled out, for the first time, the *higher intelligence-higher scholastic achievement-higher status* formula, which will become one of the core themes of the hereditarian theory of IQ: "...it is a new proof of that truth, to be held in opposition to so many paradoxically-minded persons, that the first in school are likely to be the first in life" (Binet & Simon 1911/1916, p. 291).

Binet of 1911 was different from Binet of 1905. One important point concerned his attitude towards mass testing versus individual case studies. In his case studies between 1890 and 1900, Binet observed and questioned individuals who exhibited extraordinary talent in a particular field (e.g. chess) in detail, and from there he tried to peek into their minds. The reason for examining the extremes is that phenomena (e.g. the cognitive processes underlying chess ability) would manifest themselves more clearly at the extremes. The results showed how weak any simple ranking notion was, and the rich and heterogeneous nature of human cognitive abilities.

This clinical experience would be reflected in Binet's views on how to approach the first intelligence scale. For him, the test was a diagnostic tool to distinguish between different degrees of inferior intelligence, and its clinical use should not employ automatic scoring, but active observations, interpretations, descriptions, and judgments of the investigator (Binet & Simon, 1908/1916, p. 239). However, the scale should also provide a strict threshold for diagnosis, and this required a measure of reliability: "first, repeated measurements must give essentially the same results; and second, different operators should not significantly change

the instrument's measurements" (Carson, 2007, p. 142). The element of idiosyncrasy of clinical observation would be tamed by the standardized application of the scale. This tendency was reflected in the 1908 and 1911 revisions where the tests became more and more "objective" and quantitative, and they left less room for clinical interpretation. The 1911 scale equated "normal" with "average" in a given population and allowed for rank ordering of normal individuals as well. Applicability to larger populations of schoolchildren, the need for educational as well as occupational placement based on mental capacity must have seemed attractive for Binet at last.

Testing was by no means just an intellectual exercise for Binet, it was a tool for an educational reform, which was supposed to go beyond the formal requests of the French authorities. He would like the education of schoolchildren to be organized according to their level of intelligence, and this would lead to efficient employment as well:

> Without doubts one could conceive many applications of the process, in dreaming of a future where the social sphere would be better organized than ours; where everyone would work according to his known aptitudes in such a way that no particle of psychic force should be lost for society (Binet & Simon, 1908/1916, p. 262).

One can find, in an embryonic form, almost all the central themes of the history of IQ testing in Binet's work, except for the hereditarian element that would dominate the field for a century.[7]

2. IQ Testing in the US: Standardized Tests, Standardized Intelligence

Binet-Simon scale was translated into English by the American psychologist Henry Herbert Goddard. Goddard, who was a disbeliever of Binet scale at first, was converted when he realized that the test was giving results compatible with the assessment of his assistants in the

[7] **Binet's work is incomparable to the American version of intelligence research in many respects. Binet's contribution to psychology is diverse, ranging from experiments on multiple personality disorder patients (Binet, 1896) to philosophical discussions on the mind-body problem (Binet, 1907), from the psychology of microorganisms (Binet, 1903) to the memory processes of expert chess players (Binet, 1966). The only comparison being made here is about the Binet of intelligence tests and the only point shown is that one can find ideas very similar to the American testing movement in his papers between 1905-1911.**

New Jersey Training School for Backward and Feeble-minded Children at Vineland (Kevles, 1986). Although Goddard's original translation of the 1908 version was useful in his studies on feeble-mindedness, the test had to be adapted to the American culture and should be standardized for mass administration. This was done by Lewis Madison Terman.

Terman adapted and standardized Binet's test and transformed the score to a quotient, which was calculated as the mental age divided by chronological age and multiplied by 100. Beginning in 1910, he and Childs tested almost 2300 children of different ages and took the data from 905 for statistical analysis (Terman, 1916a). The standardization sample consisted of children from a California school district of average social status (most were middle-class white children) and the data from foreign-born children were removed. He standardized the administration of the tests, changed the locations of certain tests, and added 40 items to replace the unsatisfactory elements in the Binet scale. The procedure was a direct descendant of Binet's idea: the main goal was to design tests for each age group such that the average child in that group would score 100 points. To achieve this, he had to remove questions that were too difficult or too easy, producing a test with average overall difficulty. The end-result was the first version of Stanford-Binet scale.

The scale was the first of its kind, being administered to a representative and fairly large sample. But there were problems to be resolved, such as the varying magnitude of age grades. For instance, in younger ages, mental age differences corresponded to higher magnitudes of IQ differences than older ages, which Terman called "shrinkage of mental age units" (Terman and Merrill, 1937, p.26). If the test was going to be used for ranking, this did not cause a severe problem because deviations from the mean, not the exact values, were important. However, he continued to use IQ scores for pragmatic reason, recognizing test administrators were not well-versed in statistics.

The distribution of the scores in 1916 standardization was almost “symmetrical”, approximating the normal curve (Terman, 1916a, p. 66). From this, he concluded that intelligence shows continuous distribution with no cutting line between different classes of “brightness” or “dullness”. In addition to this, the distributions for each age group were similar, thus he believed the test was measuring a fairly stable property. The almost normal distribution was the result of a pilot selection of items and the changes made in the scores. The same method was used in the 1937 revision as well. This time, Terman understood that normal distribution was not a natural phenomenon, that the test did not “measure intelligence as linear distance is measured by the equal units of a foot-rule” (Terman and Merrill, 1937, p. 25). Criticizing Thorndike (1926), who tried to equate intelligence with a linear biological variable (i.e. number of possible neural connections), he emphasized the inevitability of bootstrapping at that immature stage of their science:

> Secondly, the only available statistical procedure for making an equal-unit scale rests on the assumption that in an unselected population the distribution of intelligence follows strictly a normal curve. This may or may not be true. There are biological characters for which it is not true, and intelligence may conceivably be one of them. The question could be answered for intelligence if we had an equal-unit scale to begin with, but we are in the unfortunate position of having to assume the answer in advance, in order to derive the equal-unit scale. It is the old problem of lifting oneself over the fence by one’s bootstraps” (Terman and Merrill, 1937, pp. 25-26).

Bootstrapping was exactly what they did. Test items were ordered to ensure that the passing percent was 50% for each mental age group. The number of items also increased. Including questions from all levels of difficulty and increasing the number of questions was key to obtaining the normal distribution. As McNemar (1942) has observed, test item difficulty could be set to arrive at many different population-level distributions of “intelligence” and Terman’s selection of questions was the main reason why the scores were normally distributed.

The third problem was the small average difference between the scores of females and males. Females before age 14 were on average doing better than males and the pattern was reversed afterwards (Terman, 1916a). Terman at first believed it was insignificant, but in the 1937 revision, he and Maud A. Merrill, while ensuring the normal distribution for age groups, also handled this minor discrepancy: “This was done for the sexes separately as a basis for eliminating tests which were relatively less ‘fair’ to one sex than the other” (Terman and Merrill, 1937, p.22).

Terman (1916b) used the test as definitive of “feeble-mindedness”. The old commonsensical definition was based on one’s proper functioning in society, as expected from every normal member of society. Society’s criteria were overly dependent on historical or cultural contingencies. Thus, he decided to define the trait with a certain test score, which corresponded to 11 years of mental age in adults (i.e. 68 IQ points). His validity evidence was intuitive and reflected middle class values in the US: skilled professionals were above that most of the time and unskilled workers showed a greater ratio below it. Three years later, he would refer to the high correlations between the same test administered by different examiners as strong evidence for the validity of the scale (Terman, 1919, p.142). Then, in the 1937 revision, validity of the test was evaluated by how well it aligned with the original SB. Validation in early test construction was a circular process.

The inadequacies of the Stanford-Binet scale encouraged David Wechsler to develop his own scale. David Wechsler, as the chief psychologist in Bellevue Hospital, had been using intelligence tests for diagnostic purposes before developing his first scale, the Wechsler-Bellevue Scale, which is the ancestor of all WAIS and WISC versions. One severe problem with SB for adolescents and adults was that it was heavily weighted for verbal items. People experiencing difficulties in articulating their thoughts by language would not be assessed properly. People with a non-English native language, people with social anxiety, or people

coming from cultural backgrounds where scholastic achievement is undervalued, would be falsely diagnosed with mental deficiency (Frank, 1983, p. 7). This problem was supposedly solved later with the non-verbal Army Beta test. What Wechsler originally did (in his 1939 version) was to take Army tests, add Digits tests to the verbal subtests, combine this into a unitary scale and get rid of the mental age idea.

Wechsler's work was a continuation of the bootstrapping process, as his improvements would show. In SB, IQ was assumed to be constant across ages. This was true for the average of each age group in the standardization samples, but the variability significantly changed by age. A person 2 SD below average would have 76 points at age 6, 81 at age 10 and 84 at age 14 (Wechsler, 1941, p. 26). Wechsler believed that this discrepancy was an expected result given that mental development does not follow a linear path and that IQ formula assumed a linear growth curve in childhood. Adult intelligence posed another problem. Firstly, scales based on mental age used varying end points (14-18 ages) where development is supposed to halt, and they ignored the possibility of any negative or positive development after that. To keep IQ constant, he transformed raw scores into standard deviation units, hence IQ was no more a quotient as the name suggests. The revision was made on pragmatic and somewhat arbitrary grounds such that, the reference point was set to one probable error below, for the sake of statistical convenience and 90 points was assigned to that point because it was close to the customary 100 points (Wechsler, 1941, p. 34). By these changes, the nuisance of development, learning and deterioration of intelligence has been eliminated. IQ retained its original function of rank ordering individuals, as a score designed to be constant "throughout the life of an individual" (p.35). The classification scheme arising from these revisions was also based on a handful of "conventional" assumptions such that the average of various estimates of the percentage of mental deficiency is a good estimate for the cutting point, that the distribution of above and below average intelligence classes are symmetrical, etc. The

classification was tailored fit for a normal distribution, where the successive intervals were in units of probable error. Even after the revisions, the 1958 standardization results were not normally distributed, rather, they were skewed towards the lower end. Wechsler estimated the normal curve as the one that best fits the actual data: “The distribution of the IQ's, however, is not truly Gaussian. A curve fitted to the data would more nearly approximate Pearson's Type IV, but *the difference is not sufficiently great to be of practical significance*” (Wechsler, 1958, pp. 107-108, emphasis added).

Wechsler (1958) was critical of the circular validation in the Stanford-Binet revisions. One sort of evidence was high correlation with previous tests, but this required one to trust in the validity of the older test. He was also reminding Terman that Binet had developed the scale to avoid the subjective assessments of teachers and asking him how school progress be used as validity evidence if the test was created to be a better assessment tool. The same was true for other types of expert judgment, such as officer evaluations in the army or manager evaluations in businesses. However, despite his criticisms, he was using the same validity criteria (Wechsler, 1939, p. 78; Wechsler, 1958, p. 108).

Circularity was a common problem for test development, and this could not be resolved within the research tradition because the external criteria by which the tests were judged were not independent of the tests themselves. Scholastic achievement, as measured by school grades, was similar to IQ scores because the contents of tests and exams were similar.[8] Another reason for the circularity was that the definition of intelligence was tracking academic intelligence from the beginning. Binet himself was reluctant to give a strict definition of intelligence at an early stage of inquiry. However, he had a definition in his mind, which was also shared by Terman. By intelligence, both referred to a high-level mental

[8] **Other criteria, such as occupational status, already tracks educational attainment, which depends on scholastic performance to a significant degree (Mackintosh, 1998). The correlation between test scores and actual job performance is lower, around 0.2 (Mackintosh, 1998; Richardson, 2000).**

capacity, which shows up in tasks that require abstract thinking and judging. These tasks were scholastic from the beginning. Thus, scholastic performance, which mostly depended on verbal abilities, was imposed on the structure of tests.

3. Definition, Measurement and Psychometric Theorizing

Definitions of core concepts in a field of research become well-founded when the field matures to a certain degree. For that to happen, measuring instruments and theories are expected to coevolve such that the defined construct becomes tractable under varying conditions, by various instruments and variations are at least potentially explainable within the theory. The theory defines the concept by articulating its relations to other concepts, by explaining the variations in the measurements, and how operationalizations work. This is an interpretation of construct validity, as defined by Cronbach and Meehl (1955). Here, the expected result is a framework where there are law-like relations among constructs and their multiple-operationalizations. The process begins with imperfect definitions because the phenomenon has not been mapped in detail, measuring instruments have not proved to be reliable yet, and the theory is either absent or immature. Thus, it is understandable that psychometric research on intelligence did not need a consensus definition of intelligence in its initial stages. Measurement, or ranking and classifying, had been more central than definitions.

Psychometric tradition had always been flexible about the definition of intelligence. There were theories that took intelligence as a single measurable power whose variation explains the variation in test scores (i.e. Spearman's monarchic doctrine of *g*). Thurstone (1924/2013), in his first theory of intelligence, defined *it* as abstract thinking. Then, after developing his own method of multiple factor analysis, he defined *them* as "… correlated multiple factors, which are interpreted as distinguishable causal functions" (Thurstone, 1947, p. 439).

Wechsler, in the first edition of his scale, under the influence of Spearman, claimed that intelligence is a type of biological energy that explains the performances in a variety of tests, or in general, intellectual tasks (Wechsler, 1941, p. 11). Then in the 1958 edition, Wechsler was convinced that intelligence was only metaphorically comparable to "energy" and that it could not be seen as a tangible material entity. He described intelligence as an abstract construct which manifests itself through "learning, adapting, reasoning and other forms of goal directed behavior" (Wechsler, 1958, p. 5). As Spearman (1961) had observed, it was not difficult to arrive at a common definition as long as everyone was free to interpret it in their own manner. But operationalizing the definition was not that easy. Boring (1923/1961) suggested confining the meaning of psychometric intelligence to "what the tests test", because of the difficulty in measuring other aspects of the everyday concept of intelligence, not probed by the tests. The paper was not written to satirize the lack of an agreed definition of intelligence, but rather to ensure that the definition should be limited to those traits that tests can test. One consequence of this is to exclude from the definition of intelligence those abilities that tests cannot test, such as specific aptitudes. For example, according to one criticism of intelligence tests, placing too much emphasis on speed is to the detriment of people who move at a slow but sure pace and solve problems in a precise manner. Boring (1923/1961) responded to this criticism with the following example:

> If these people have less power, they have to go up the hill on low gear and it takes them longer; that is all. Of course they 'get there' just the same, but when they 'get there' their powerful rivals are on and somewhere else. If they ride more smoothly as they go, that is an entirely different matter from the one under discussion; they have a special ability which is not intelligence as the tests test it (p. 212).

Boring (1923/1961) was emphasizing that the colloquial concept of intelligence (i.e. problem-solving ability) was not the same as psychometric intelligence, and that intelligence in psychometrics was an abstraction made from the relative constancy of intelligence test

scores (i.e., constancy of rankings) and the imperfect but significant correlation between subtests. If intelligence is taken as people's ability to solve problems, it would improve with each new specific skill learned, and this would make it extremely complex to be assessed on a single scale. The solution was to define intelligence solely by one of its possible operationalizations, that is, by tests. However, the interpretation of test scores was going way beyond this restricted operationalism, and tests were seen as measuring some real psychic variable. This is what Block and Dworkin (1974) meant when they said that psychometricians want to "...have their cake and eat it too" (p. 355).

The contrast between physical measurement and IQ testing reveals the tensions between Boring style operationalism and realism concerning IQ. Physical measurement was based on the classical theory which says that measurement is finding out the ratio of the magnitude of a quantitative attribute to the unit of the same attribute. In this theory, for an attribute to be measurable, it must have a quantitative structure, in short, it must be additive (Markus and Borsboom, 2013). Additivity need not be interpreted as concatenation of physical magnitudes as happens in the case of length, but as a magnitude being exclusively composed of discrete magnitudes (Michell, 1997, p. 357).[9] This type of compositional structure is missing in test scores, it is not possible to decompose an IQ score to its components in this manner. When intelligence researchers realized that the strict rules of measurement in physics did not fit their field, they again resorted to a circular route, and they developed a concept suitable for their practice. Operationalism was the first option because of its lower standards, and representational measurement theory developed out of it allowed for any numerical assignment based on a well-defined procedure to count as measurement (Markus and Borsboom, 2013, p.27). When this turned out to be overly liberal, the axiomatic version of the representational theory, which constrained measurement by adding an element of

[9]**We might call this operation "combination" rather than addition, as Bostock (1979, p. 104) suggests.**

isomorphism between the numerical assignment and an empirical structure, was employed (Markus and Borsboom, 2013, p. 32). In axiomatic theory, the structure of numbering should be projectible onto an empirical structure, thus, not every conventional procedure of numbering would count as measurement. In the case of IQ, what is the empirical structure to compare with test score distributions? Teacher's assessments, social status, or other intuitive benchmarks do not seem to be better than test scores. Given that neither the unconstrained representational nor the axiomatic versions of operationalism worked with an attribute like intelligence, psychometric research on IQ has moved from the observable to the unobservable. The latent factors extracted in factor analytic theories began to be seen as the dimensions of intelligence indirectly measured by the tests.[10]

The fundamental assumption in latent factor theories is that test scores are *caused* by the values of and the relationships between certain unobservable variables. For instance, in Spearman style monarchic theory, an individual's level of *g* would determine the "true score", to the degree of *g-loading* of the test. Latent factors are inferred from the correlations between test scores and they are assumed to be the common cause that explains subtest correlations. When the latent variable is controlled for (i.e. sample is stratified according to latent factor values), that would lead to local independence between previously correlated scores (Bartholomew, 2004; Markus and Borsboom, 2013). One advantage of latent factor theories is that they *seem to* break the semantic equivalence between measurement and the attribute measured, hence, they can avoid a tautological version of operationalism. However, the attributes or *abilities* are dependent on the test contents, and this creates problems for construct validity.

[10] **Here I omitted the well-known indeterminacy problem about factor analysis, that is, it cannot distinguish between completely different underlying causal structures if they produce the same distribution of correlations, and focused on the circularity of the entire process (Gould, 1996; Clapp Sullivan et al., 2024). Although an unresolved and crucial problem for factor analysis, it is not central to the argument in this paper.**

Psychometric models derived by factor analysis are content neutral in the sense that the models show just how to organize correlations into clusters. In the case of intelligence research, this means the models themselves do not have labels on their nodes that say, “I am verbal ability” or “I am visuospatial ability”. The attributes in a correlation table, or the factors extracted from them do not have any functional significance or meaning by themselves. It is the researchers who provide the meaning “...from our prior insight into the composition of mental abilities and traits” (Herrnstein, 1973, p. 92). Intuitive-verbal theories and expert opinions, which are mostly based on test item contents, are the main sources of labeling, and these are far from providing a principled method of deciding on what is really being measured. For example, in one of the founding texts of the Cattell-Horn-Carroll (CHC) theory, cognitive abilities (the latent factors being measured) “are differentiated not only by the fact that their intercorrelations are often far from perfect, but also by the fact that they pertain to different classes of tasks” (Carroll, 1993, p. 712). Task here corresponds to answering test questions and there was no theory connecting this with the cognitive processes involved. Almost twenty years ago, Carroll (1976) was proposing a method “….to *start* from a theory of cognitive processes and *then*, on this basis, to attempt to characterize FA factors and, by implication, what the corresponding FA tests measure” (p. 30). In his *opus magnum*, he seemed to have abandoned this project.

Factor analytic theories, especially of the monarchic theory of Spearman, have been heavily criticized for reifying a statistical artifact (Gould, 1996). Carroll responded to this criticism by interpreting factors as intervening variables rather than hypothetical constructs (i.e. unobservable entities or powers that would one day be observable). However, he was still considering cognitive abilities (i.e. group factors) as if they were *internal potentialities of individuals* that explain their scores in a certain test: “It is the underlying *ability* that is

relevant to success in business administration, not the knowledge of particular items" (Carroll, 1993, p. 24).

Exploratory factor analysis depends on certain assumptions such as the test scores measure a set of abilities that "act as a functional or operational unit" (Carroll, 1993). More specifically, test performance is taken as if it results from a person's "strength" in each factor, multiplied by the weight of those factors in the task (the factor loading) and summed up. The basic idea is that performance is a linear function of one's cognitive abilities and the relevance of those abilities for the test. The latent abilities are assumed to be normally distributed and the essential constraint on the number and relations of factors (e.g. factor loadings) is that they reproduce the correlations of test scores (pp. 50-51). Thus, what tests are included in the study affects what factors would be extracted. In this regard, factor analysis itself is hypothesis free with respect to the resultant factor structure, but the theories so derived will be classificatory rather than explanatory, regardless of the underlying assumptions[11]. In short, there is no way to access those abilities other than the test scores and their covariance.

One possible solution to break this circle is to find an independent measure of intelligence, not contaminated by social and cultural biases and the circularity imposed by intuitive definitions, as scholastic achievement or "occupational prestige" do. A reasonable place to start searching were the cognitive processes and their neural correlates. There have been certain brain-based theories of intelligence, which identify certain neural correlates of intelligence, such as the energy consumption of the brain in carrying out cognitive tasks, brain size, response times and nerve conduction velocity (Bartholomew, 2004, p. 53; Jensen quoted in Miele, 2002/2019, p. 63). The functional significance – meaning – of these

[11] **For instance, Bartholomew (2004), while discussing the difference between factor analysis and principal component analysis (PCA), correctly states that PCA is a dimension reduction method with no causal or ontological claim, but factor analysis produces a latent causal structure to explain the observed correlations. However, this is just an assumption and the factors so derived seem not much better than principal components with respect to their explanatory power.**

correlates would still depend on how well they align with the traditional measures of intelligence, like test scores (Fletcher & Hattie, 2011, p. 28). In other words, if the trait is still described within the conceptual framework of psychometrics, how "deep" one goes will not make a difference. To break this circle, something more than better standardization, culture-fair norming or other psychometric virtues is needed. Robustness is one of the missing elements.

Robustness is "the use of multiple independent means to detect, derive measure, manipulate, or otherwise to access entities, phenomena, theorems, properties and other things we wish to study" (Wimsatt, 2007, p.37). Multiple, independent measures are essential for verifying the existence of the property or entity or for testing the validity of a model. This point has been recognized in contemporary psychometric research. New batteries come with validity evidence from multiple sources, such as external criterion correlations, factorial structure (i.e. whether they measure CHC abilities in different samples), correlations with other tests, etc. Sometimes, the existence of multiple models or instruments of measurement does not guarantee robustness, because they might not be truly independent due to shared assumptions, similar deep structures, or common methodological biases (Wimsatt, 2007, p.72). This seems to be case for IQ tests. External validity criteria for IQ tests include scholastic achievement (e.g. school grades), not an independent measure.[12] Factor structure itself is derived from test score correlations, thus not independent. Two tests having the same factorial structure show nothing other than the content similarity.

Another crucial aspect of robustness is the semantic (or theoretical) irrelevance (or independence) of the intermediate steps in measurement from the attribute being measured.

[12] **Evaluation of validity has become a context dependent and pragmatic issue, as Han (2024) observes. It is a measure of how strong a certain interpretation of a test is backed by evidence. A stronger concept of validity, such as Stone's (2019) idea that a measure tracking the variations in a construct is more attractive, but seems unrealistic in intelligence research due to the multidimensionality of the phenomenon measured.**

The principles of the workings of the measuring instruments build a causal chain from the variations in the thing being measured to what is happening inside the instrument and finally the observed results. The causal steps in this chain should not have a definitional bearing on the attribute being measured. For instance, temperature is not defined by what happens to a mercury thermometer when it is immersed in boiling water. But in the case of intelligence research, the instrument itself is definitive of the measured entity. Latent factors, as derived from the correlations between test scores, provide a detachment of the instrument from the ability they are supposed to measure. However, the nature of these factors is not known and there is no satisfactory theory of how they influence test scores. Factorial theories do not provide a strong ground for judging the validity of tests (Markus and Borsboom, 2013, p. 279).[13] More and more correlational studies to provide evidence for the enterprise[14] do not compensate for the lack of genuine progress. As Thurstone (1937, Preface) had warned almost 90 years ago:

> What is needed in experimental psychology more than anything else is to formulate problems and investigations so as to reveal functional relations which should be rationalized whenever possible. This will advance psychology toward scientific respectability with more certainty than correlation coefficients, elaborate instrumentation, and discussion about points of view and the meaning of words.

The investment in the measurement tool was spared from theory construction. The theories produced by factor analysis were based solely on the score variance overlaps, which depend on the contents of tests. After the CHC theory became the dominant framework, new tests began to be judged according to how good they measure the abilities (especially the broad abilities) posited there, i.e. it became the benchmark in test construction. For example, when Kaufman developed his K-ABC, which was based on Sperry's (1968) brain asymmetry

[13] **I do not share Haig's (2005; 2018) optimism about exploratory factor analysis as an abductive method of discovery, because of the over-reliance on test score distributions.**

[14] **For example, Schneider, W. J., & McGrew, K. S. (2018) and Flanagan & Dixon, (2014) provide ample correlational evidence, ranging from academic outcomes to neuropsychological assessments, without any explanation of why these correlations hold.**

theory and Luria's (1966) sequential-simultaneous information processing theory, it was criticized for not measuring broad abilities in CHC (Kaufman, 2009, p. 70). CHC revisions, as Schneider and McGrew (2018) describe, are incremental and adapt to empirical findings rather than anticipating them. These points are symptomatic of a degenerate bootstrapping process.

4. Parieto-Frontal Integration Theory: Brain imaging in and out of the psychometric box

In the previous section, I argued that the diagnosis of "degenerate bootstrapping" can also be applied to brain-based intelligence research because the interpretation of these studies is driven by psychometric assumptions. I will elaborate on this point in the context of a recent brain-based theory of intelligence – the Parieto-Frontal Integration Theory. We can summarize the result as follows: The P-FIT-based research program fits the degenerate bootstrapping diagnosis insofar as it remains dependent on categories of cognitive ability derived from psychometric measures of intelligence and factor analysis, but it is progressive insofar as it contributes to what we know about the anatomical structure of the brain and the neural substrates of cognition. The same is true for alternative frameworks such as the chronometric neural efficiency theory of Jensen (2006). In other words, brain-based theories serve scientific progress to the extent that they step outside the psychometric box. Of course, a few caveats about "progress" are in order here. In Lakatos' theory, whether a research programme is progressive or not is understood over a period of time, the length of which is not clearly defined. Considering that P-FIT was introduced in 2007, it might be said that the time that has elapsed since then is too short to make such an assessment.

A second caveat concerns the difference between physics, where Lakatos developed his criteria for progress, and intelligence research. In a relatively new field such as neuroscience, in the study of a structurally and functionally complex organ such as the brain, and in the

study of complex psychological phenomena such as human cognitive abilities, it is difficult to expect Lakatos' criterion of “continuous theory-based growth” to be met. Lakatos' unpleasant diagnosis of the use of statistical methods in the social sciences mentioned in the introduction reflects a very strong theory-based understanding of science. In this understanding, the minimum condition for being a research programme is to have a hard core that is resistant to revision, and progress requires a research practice that extends to new phenomena while preserving this core. I think such a rigid understanding of progress would not take us far in a fairly recent field, and interesting observations alone can serve progress, if they inspire a bootstrapping process of observation-theory construction-novel predictions. In this sense, I think P-FIT and related “brain efficiency” theories can contribute to scientific progress, regardless of the psychometric elements (e.g., *g*) in their theoretical foundations. The reason is that, whether the theoreticians aim it or not, their findings have a tendency to break out of the psychometric box and call for a more radical theory revision, or for the construction of a new theoretical foundation.

P-FIT can be described as a localized “brain efficiency” theory. Localization here should be broadly understood as identifying the cortical areas that form a brain-wide network and not pinpointing a specific homunculus in the brain. Brain efficiency theories are based on the intuition that the speed and accuracy of information processing in the brain is the key to explaining intelligence differences, and this depends on the overall structure of the brain. For instance, Jensen (2006) states that processing speed is an important determinant of how fast people solve problems, how efficiently they integrate information, how fast they learn, how accurately they assess information (e.g. prioritize evidence) and come to decisions, and thus, how intelligent they are. Here, intelligence refers to *g* and according to Jensen, it is not localized in any specific part of the brain. Jensen's chronometric theory focuses on the objective measurement of intelligence and aims to transform it into a ratio level variable,

rather than uncovering the biological basis of it. P-FIT, on the other hand, goes beyond measurement and aims at localizing and understanding intelligence.

P-FIT was first formulated in a paper published in 2007 (Jung and Haier, 2007). The rationale for the theory was twofold: to integrate findings that relate intelligence with brain-based measures and to act as a framework to formulate testable hypotheses. The fundamental question the theory aimed to answer was "where individual differences in intelligence might arise in the human brain" (Jung and Haier, 2007, p. 138). In summary, the theory asserts that 14 Brodmann areas mostly on the parietal and frontal cortex, along with their connections, are the seat of *g* variability among individuals. The theory describes the steps in the integration process as follows: The process begins with the reception and initial processing of sensory information in the occipital and temporal areas, then the information is integrated and abstracted in the parietal and temporal areas, and the integrated information goes to frontal areas for hypothesis testing and decision making (Haier et al., p. 132). The empirical findings that inspired the theory were diverse, brain imaging data being the focal element. In this section, I will first summarize the empirical underpinnings of the theory, then point out to the inconsistencies and reveal the assumptions it inherits from psychometrics, and finally evaluate the relationship between the empirical findings and these assumptions through the concept of degenerate bootstrapping.

What set of evidence led to the formulation of P-FIT? First of all, P-FIT tracks variability in *g* and not specific abilities. In a sense, the research programme around P-FIT focuses on the neural correlates of *g*. One important finding that connects IQ tests and brain function was the discovery that PET (positron emission tomography) images showed less brain activity in people with high IQ scores (Haier, 2023). The first such imaging study by Haier (1988) peaked into the brains of participants while they were solving Raven's Progressive Matrix problems. Brain activity and IQ was negatively correlated. This result led to one version of

the “brain efficiency” hypothesis: “higher intelligence requires less brainwork” (Haier, 2023, p. 80). Comparable results in other cognitive tasks (e.g. verbal tasks) were obtained by other researchers. Studies that compared the brain activity in naive vs expert Tetris players also corroborated this picture.

Structural variables, especially the organization of gray and white matter, were also correlated with *g.* One such finding was that “g accounted for many of the FSIQ correlations with gray matter” in various areas of the brain, including the parietal and frontal areas (Haier, 2023, p. 96). *g*-loadings of each subtest were predictive of how strongly the scores correlated with gray matter. Another type of imaging study provided evidence that white matter density and organization in parietal and frontal areas correlated highly with IQ scores. This finding was important because white matter is what connects various brain regions and is vital for the integration of information from different neural circuits. Graph analysis of fMRI data revealed that the length of the path between parietal to frontal areas was negatively correlated with IQ, the shorter the path, the higher the IQ.

Higher cortical thickness and lower dendritic branching were also shown to be associated with higher levels of intelligence (Genç et al, 2018). In another study, high dendritic complexity in pyramidal neurons was associated with intelligence. The emerging picture is that a thicker cortex with less dense neurites and larger dendrites in the P-FIT network are positively correlated with high IQ scores (Haier et al., 2023, p.148). High IQ individuals also show less variability in the connectivity patterns in fMRI studies of resting state brain activity. Other studies that probe network reconfiguration in key brain areas with changing tasks found that less reconfiguration occurred in high-*g* individuals (Haier, 2023, p. 116).

These findings, taken together, suggest that intelligence as the tests test it, is related to an interactive process of learning and brain maturation. It is about the differences in how the

developing brain is shaped by learning in childhood and adolescence. This would also explain the similarity of results in the Tetris study and IQ-PET studies mentioned above. This further suggests that high IQ is related to "been there, done that" kind of brain structure. In other words, the high-IQ brain is one that has settled on certain solutions on a range of problems it has experienced, and it can generalize these solutions via filtering out irrelevant processes and inputs. These remarks are tentative because heterogeneity of the neural correlates of intelligence complicates the development of a general theory of intelligence based on imaging studies. People can have the same IQ scores or *g* with varying brain sizes, cortical thickness, gray matter density or neural network activity patterns.

P-FIT has been facing this heterogeneity challenge from the initial moments of its development. One of the first findings that seemed to contradict the brain efficiency hypothesis was that males with higher math skills showed higher brain activity when solving SAT math questions, whereas there was no significant correlation between brain activity and ability in females (Haier, 2023, p. 85). The brain efficiency findings mentioned above are also inconsistent with newer studies which also cover task difficulty as an additional variable. Basten et al. (2013) study found that when the task is more difficult, task specific networks show greater activation (less efficiency) and less activity in other networks in high-IQ individuals. Another inconsistency concerned cortical thickness-intelligence correlation. In a longitudinal study of children and adolescents, it was found that children with higher IQ scores had thinner cortices early on, then their cortices thickened for a longer time compared to others, and they experienced a substantial cortical thinning at early adolescence (Shaw et al., 2006). Age, sex and socioeconomic status are among the factors associated with heterogeneity.

P-FIT founders recognized and even embraced this heterogeneity. When the theory was first formulated, it was apparent that P-FIT was not a one size fits all type of theory. Jung and

Haier (2007) were open to revising the theory with new observations: "The P-FIT will evolve as we follow these new observations" (p. 175). Later, Haier (2023), formulated three rules of thumb in interpreting the findings of neuroimaging studies of intelligence, which I call the principles of humility: "No story about the brain is simple, no one study is definitive, and it takes many years to sort out conflicting and inconsistent findings and establish a weight of evidence." (p. 86). These remarks suggest that the theory, apart from pointing out to specific brain areas as the seat of intelligence, provides a fairly flexible framework for the interpretation of those findings. This flexibility is not achieved by positive heuristics resolving anomalies, as would be expected from a Lakatosian progressive programme. Rather, the hard core itself seems to be nonexistent. The theory does inherit its foundational principles from classical psychometrics. The assumptions inherited include a realist interpretation of psychometric *g*, belief in a strong genetic basis of individual differences, and the classical validity criteria for IQ tests. The problem of causal interpretation of correlational imaging data aside, these assumptions create the same circular process as imaging findings are judged according to how well they conform to psychometric benchmarks.[15] Intelligence is equated with *g,* which is supposed to be measured by full scale IQ.[16] Then IQ differences are projected onto structural and functional variation in the brain. For instance, brain-based (e.g. gray matter size, speed of information processing) metrics correlated with psychometric intelligence are supposed to be used for extracting a *brain-based g* (Haier, 2023). The question of validity repeats itself here: what is the relation of this new measure to

---

[15] **One interesting suggestion in the literature is that psychometric *g* is useful as a bridge model between folk psychological intelligence and cognitive and neural underpinnings of intelligence (Curry, 2021). It acts as a double-sided filter that abstracts away irrelevant details of folk intelligence and neural mechanisms, creating a good fit between the two domains. I agree with this function of *g* but I believe this reciprocal filtering and narrowing works against novelty, further solidifying the circularity in intelligence research.**

[16] **Another problematic assumption is that IQ scores are on an interval scale (Haier, 2023, p. 187). As shown in the third section, this is an artificial consequence of test design, norming and interval scale is assumed for the ease of statistical analysis.**

intelligence? Correlation with IQ, predictive validity criteria, etc would be used to answer this question. There are no explicit bridge principles that connect these structural and functional brain variables to cognitive processing and further to general intelligence, and absent these principles, the entire process seems to repeat the circularity on a different level. Heritability of IQ, along with the findings of newer genomic association studies are assumed to provide further support. Brain development is supposed to be strongly controlled by the genome, minimizing the effects of learning. The Fisherian polygenic and additive model of inheritance, which was historically employed to explain continuous and normal distribution of IQ, also grounds the continuous variation of brain based measures (e.g. processing speed, neurite density, etc.).[17] "Graded genes, graded brains, graded intelligence" type of model seems to be at the core of P-FIT (Richardson, 2022). Model fitting practices imposing a structure to a representation of nature (e.g. brain variation), and then, nature so represented being used to support the practice perpetuates the circularity to new levels of inquiry.

Once again, we observe bootstrapping in practice. This time, psychometric *g* acts as the benchmark for judging imaging studies and imaging studies are invoked to support the psychometric vision. Rather than looking thoroughly how the implicated brain areas function when individuals perform tasks, research focuses on how well anatomic or physiological variables correlate with *g.* Differential psychology's attempt to confine the field to a psychometric box leaves behind important questions such as how learning takes place, how the acquisition of expertise is reflected in the brain, how brain adapts to the developmental

[17] **Fisher's (1918) model was proposed to explain the familial correlations of metric characters under Mendelian assumptions. Fisher proved that infinitely many Mendelian genes of infinitesimal and additive effects would explain the continuous distribution of such traits. Cyril Burt was one of the first intelligence researchers to apply this model to the inheritance of intelligence, where uncorrelated, small and cumulative genetic effects were assumed to create the normal distribution (Burt and Howard, 1956, p. 97). Jensen (1969) also based his notorious hereditarian argument on this model. Modern genome-wide association studies (GWAS) employ similar assumptions, ignoring factors that might complicate the causal effects of genes. However, detecting gene-gene (epistasis) and gene-environment interactions from GWAS data is not as easy as detecting trait associated genetic variants (Balvert et al., 2024). The methods are more suited to detecting additive effects, and this creates a bias in results.**

environment, how culture is internalized, how cultural differences between groups affect problem solving methods and brain structures underlying them.[18] Despite these limitations, the field is ripe for progress. The data generated by those studies has the potential to help identify the processes and structures grounding human intelligence, regardless of the circular conceptual structure of psychometrics that still guides such research.

Conclusion

Intelligence testing had been one of the most controversial areas of psychology in the 20th century. The methodological flaws, beginning from definitional issues and extending towards the status of factor analytic theories, have been extensively debated. It has been almost 30 years after the storm created by the *Bell Curve* and IQ testing seems to have become less and less relevant to the lives of the majority of the younger generation. Here, I tried to articulate the main methodological problems in the field, concerning the definition and measurement of intelligence, validity inferences and factor-analytic theory construction. A unifying theme has been degenerate bootstrapping – a degenerate form of auto-refinement process – which led to the less-than-optimal scientific performance of the field.

Psychometric theorization on intelligence was placing too much emphasis on test scores and their correlations. The overlap of variance was explained by a common cause model where factors as latent variables were responsible for the observed distribution of scores. The problem with extreme operationalism in the testing community had found its remedy in these theories. However, these theories extracted latent variables from test score correlations and even the naming of the factors was nothing more than intuitive and conventional. In other words, factors did not correspond to some deeper cognitive process that can be identified and studied. This was one of the reasons why the field did not pass the robustness test. Reification

[18]**Almost all of the studies Haier (2023) mentions as examples of progress (e.g. neuroimaging studies of mnemonic genuises, single-neuron studies in animals, etc) have little in common with psychometric intelligence research.**

is not a problem by itself, but once one reifies a factor, then one must find multiple ways to access it.

The limitations of the validity of a construct can reflect the limitations of evidence, and new evidence can be a basis for the revision of the test as well as the psychological theory of the construct. Thus, some kind of bootstrapping was and still is inevitable in intelligence research. The problem with the testing movement and the psychometric theorizing was that they have built a closed circle of reasoning that prevents other types of information (from cognitive neuroscience, evolutionary theory, animal behavior research, anthropological studies on cognition, etc.) to be integrated into their framework to build a better theoretical foundation. One brain-based theory (P-FIT) expanded the horizons of intelligence research by investigating the neural underpinnings of intelligence differences. However, due to its reliance on the psychometric conceptual framework, it came short on providing a truly bottom-up theoretical foundation. Such a foundation would provide constraints on the psychometric models employed and increase their empirical content, thus leading to a progressive type of bootstrapping (Borsboom, 2006). The technological developments in test construction, statistical analysis and validation have led testers to walk faster and with surer steps, but on the same circle as Binet, Terman and Wechsler walked almost a century ago. This is what degenerate bootstrapping amounts to.

References

Balvert, M., Cooper-Knock, J., Stamp, J., Byrne, R. P., Mourragui, S., van Gils, J., Benonisdottir, S., Schlüter, J., Kenna, K., Abeln, S., Iacoangeli, A., Daub, J. T., Browning, B. L., Taş, G., Hu, J., Wang, Y., Alhathli, E., Harvey, C., Pianesi, L., … Twine, N. A. (2024). Considerations in the search for epistasis. *Genome Biology*, *25*(1), Article 1. https://doi.org/10.1186/s13059-024-03427-z

Bardini, Thierry. (2000). *Bootstrapping: Douglas Engelbart, coevolution, and the origins of personal computing*. Stanford University Press.

Bartholomew, D. J. (2004). *Measuring Intelligence: Facts and Fallacies*. Cambridge University Press.

Binet, A. (1896). *On double consciousness: Experimental psychological studies* (New ed.). Open Court Pub. Co.

Binet, A. (1903). *The psychic life of micro-organisms: A study in experimental psychology.* Open Court Pub. Co.

Binet, A. (1907). *The mind and the brain: The authorised translation of L'âme et le corps.* K. Paul, Trench, Trübner & co. ltd.

Binet, A. (1966). *Mnemonic virtuosity: A study of chess players* (M. L. Simmel & S. B. Barron, Trans.). Journal Press.

Binet, A., & Simon, T. (1905a). Upon the Necessity of Establishing a Scientific Diagnosis of Inferior States of Intelligence (E. S. Kite, Trans.). In A. Binet & T. Simon (1916), *The development of intelligence in children* (pp. 9–36). Williams and Wilkins Publishing Co.

Binet, A., & Simon, T. (1905b). Application of the New Methods to the Diagnosis of the Intellectual Level among Normal and Subnormal Children in Institutions and in the Primary Schools (E. S. Kite, Trans.). In A. Binet & T. Simon (1916), *The development of intelligence in children* (pp. 91–181). Williams and Wilkins Publishing Co.

Binet, A., & Simon, T. (1908). The Development of Intelligence in the Child (E. S. Kite, Trans.). In A. Binet & T. Simon (1916), *The development of intelligence in children* (Limited ed., pp. 182–273). Williams Printing Co.

Binet, A., & Simon, T. (1911). New Investigations upon the Measure of the Intellectual Level Among School Children (E. S. Kite, Trans.). In A. Binet & T. Simon (1916), *The*

development of intelligence in children (pp. 274–329). Williams and Wilkins Publishing Co.

Block, N. J., & Dworkin, G. (1974). IQ: Heritability and Inequality, Part 1. *Philosophy & Public Affairs*, *3*(4), 331–409. https://www.jstor.org/stable/2264953

Boring, E. G. (1961). Intelligence as the Tests Test It. In J. J. Jenkins & D. G. Paterson (Eds.), *Studies in individual differences: The search for intelligence* (pp. 210–214). Appleton-Century-Crofts.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Bostock, D. (1979). *Logic and Arithmetic, volume II* (Vol. 2). Oxford University Press.

Branahl, J. (2024). *Stagnant Lakatosian Research Programmes* (arXiv:2404.18307; Version 2). arXiv. https://doi.org/10.48550/arXiv.2404.18307

Brown, J. (1992). *The definition of a profession the authority of metaphor in the history of intelligence testing, 1890-1930* (Course Book). Princeton University Press.

Burt, C., & Howard, M. (1956). The Multifactorial Theory of Inheritance and Its Application to Intelligence. *British Journal of Statistical Psychology*, *9*(2), 95–131. https://doi.org/10.1111/j.2044-8317.1956.tb00177.x

Carroll, J. B. (1976). Psychometric Tests as Cognitive Tasks: A New "Structure of Intellect." In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27–55). L. Erlbaum Associates.

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

Carson, J. (2007). *The Measure of Merit: Talents, Intelligence, and Inequality in the French and American Republics, 1750-1940*. Princeton University Press.

Cartwright, N., Hardie, J., Montuschi, E., Soleiman, M., & Thresher, A. C. (2022). *The Tangle of Science: Reliability Beyond Method, Rigour, and Objectivity*. Oxford University Press.

Castles, E. E. (2012). *Inventing Intelligence: How America Came to Worship IQ*. Bloomsbury Publishing USA.

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

Clapp Sullivan, M. L., Schwaba, T., Harden, K. P., Grotzinger, A. D., Nivard, M. G., & Tucker-Drob, E. M. (2024). Beyond the factor indeterminacy problem using genome-wide association data. *Nature Human Behaviour*, 1–14. https://doi.org/10.1038/s41562-023-01789-1

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Curry, D. S. (2021). G as Bridge Model. *Philosophy of Science*, *88*(5), 1067–1078. https://doi.org/10.1086/714879

Felegi, B. (2024). *The Impacts of Removing College Entrance Exams: Evidence from the Test-Optional Movement* (SSRN Scholarly Paper 4494844). Social Science Research Network. https://papers.ssrn.com/abstract=4494844

Fischer, C. S., Voss, K., Swidler, A., Lucas, S. R., Sánchez-Jankowski, M., Hout, M., & Fischer, C. S. (2006). *Inequality by Design Cracking the Bell Curve Myth*. Princeton University Press.

Fisher, R. A. (1918). XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, *52*(2), 399–433.

Flanagan, D. P., & Dixon, S. G. (2014). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In *Encyclopedia of Special Education*. John Wiley & Sons, Ltd.

Frank, George. (1983). *The Wechsler enterprise: An assessment of the development, structure, and use of the Wechsler tests of intelligence* (1st ed.). Pergamon Press.

Gebel, L., Velu, C., & Vidal-Puig, A. (2024). The strategy behind one of the most successful labs in the world. *Nature*, *630*(8018), 813–816. https://doi.org/10.1038/d41586-024-02085-2

Genç, E., Fraenz, C., Schlüter, C., Friedrich, P., Hossiep, R., Voelkle, M. C., Ling, J. M., Güntürkün, O., & Jung, R. E. (2018). Diffusion markers of dendritic density and arborization in gray matter predict differences in intelligence. *Nature Communications*, *9*(1), 1905. https://doi.org/10.1038/s41467-018-04268-8

Gould, S. Jay. (1996). *The mismeasure of man* (Rev. and expanded.). W.W. Norton.

Haier, R. J. (2023). *The Neuroscience of Intelligence* (Second edition.). Cambridge University Press.

Haier, R. J., Colom, R., & Hunt, E. (2023). *The Science of Human Intelligence* (2nd ed.). Cambridge University Press.

Haig, B. D. (2005). Exploratory Factor Analysis, Theory Generation, and Scientific Method. *Multivariate Behavioral Research*, *40*(3), 303–329. https://doi.org/10.1207/s15327906mbr4003_2

Haig, B. D. (2018). *Method Matters in Psychology: Essays in Applied Philosophy of Science* (1st ed. 2018.). Springer International Publishing.

Han, Y. (2024). Multiple Historic Trajectories Generate Multiplicity in the Concept of Validity. *Perspectives on Science*, *32*(4), 488–517. https://doi.org/10.1162/posc_a_00624

Herrnstein, R. J. (1973). *I.Q. in the meritocracy.* Little, Brown.

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

Howe, M. J. A. (1997). IQ in question the truth about intelligence. In *IQ in question the truth about intelligence*. SAGE.

Jacoby, R., & Glauberman, N. (Eds.). (1995). *The bell curve debate: History, documents, opinions* (1st ed.). Times Books.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, *39*(1), 1–123. https://doi.org/10.17763/haer.39.1.l3u15956627424k7

Jensen, A. Robert. (1998). *The g factor: The science of mental ability.* Praeger.

Jensen, A. Robert. (2006). *Clocking the mind: Mental chronometry and individual differences*. Elsevier.

Joseph, J., & Richardson, K. (2025). The Bell Curve at 30: A Closer Look at the Within- and Between-Group IQ Genetic Evidence. In R. M. Lerner & G. Greenberg (Eds.), *Heredity Hoax: Challenging Flawed Genetic Theories of Human Development* (pp. 433–476). Routledge.

Kaufman, A. S. (2009). *IQ testing 101*. Springer Pub. Co.

Kevles, D. J. (1986). *In the name of eugenics: Genetics and the uses of human heredity.* University of California Press.

Lakatos, I. (1968). Changes in the problem of inductive logic. In I. Lakatos (Ed.), *The problem of inductive logic* (Vol. 2, pp. 315-417). North Holland Pub. Co.

Lakatos, I. (1970). Falsification and the methodology of research programmes. In I. Lakatos & Alan. Musgrave (Eds.), *Criticism and the growth of knowledge* (Vol. 4, pp. 91–195). Cambridge University Press.

Lakatos, I. (1978). *The methodology of scientific research programmes* (J. Worrall & G. Currie, Eds.). Cambridge University Press.

Lewontin, R. C., Rose, S. P. R., & Kamin, L. J. (1984). *Not in our genes: Biology, ideology, and human nature*. Pantheon Books.

Lovell, D., & Mallinson, D. J. (2024). Pencils Down… for Good? The Expansion of Test-Optional Policy After COVID-19. *Innovative Higher Education*, *49*(1), 177–199. https://doi.org/10.1007/s10755-023-09677-2

Luria, A. R. (1966). *Human brain and psychological processes*. Harper & Row.

Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford University Press.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. Taylor & Francis Group.

Matarazzo, J. D. (1972). *Wechsler's Measurement and appraisal of adult intelligence* (5th and enl. ed. ed.). Williams & Wilkins.

Matzel, L. D. (2024). An endless cycle of ignorance is the consequence of not offering classes on IQ and human intelligence. *Intelligence*, *104*, 101827. https://doi.org/10.1016/j.intell.2024.101827

McNemar, Q. (1942). The revision of the Stanford-Binet scale, an analysis of the standardization data. In *The revision of the Stanford-Binet scale, an analysis of the standardization data*. Houghton Mifflin company.

McNemar, Q. (1964). Lost: Our intelligence? Why? *The American Psychologist*, *19*(12), 871–882. https://doi.org/10.1037/h0042008

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *British Journal of Psychology*, *88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Miele, F. (2019). *Intelligence, Race, And Genetics: Conversations With Arthur R. Jensen*. Routledge. https://doi.org/10.4324/9780429499753

Murdoch, S. (2007). *IQ: A smart history of a failed idea* (1st ed.). J. Wiley and Sons.

Nicolas, S., Coubart, A., & Lubart, T. (2014). The program of individual psychology (1895-1896) by Alfred Binet and Victor Henri. *L'Année Psychologique*, *114*(1), 5–60. https://doi.org/10.3917/anpsy.141.0005

Richardson, K. (2000). *The making of intelligence*. Columbia University Press.

Richardson, K. (2022). *Understanding intelligence*. Cambridge University Press.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell– Horn– Carroll Theory of Cognitive Abilities. In D. P. Flanagan, E. M. McDonough, & A. S. Kaufman (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 73–163). Guilford Publications. http://ebookcentral.proquest.com/lib/utxa/detail.action?docID=5485039

Spearman, C. (1961). The Abilities of Man. In J. J. Jenkins & D. G. Paterson (Eds.), *Studies in individual differences: The search for intelligence* (pp. 241–266). Appleton-Century-Crofts.

Sperry, R. W. (1968). *Hemisphere deconnection and unity in conscious awareness. American Psychologist*, *23(10)*, 723–733. https://doi.org/10.1037/h0027161

Staub, M. E. (2018). *The mismeasure of minds: Debating race and intelligence between Brown and The Bell Curve*. The University of North Carolina Press.

Stone, C. (2019). A Defense and Definition of Construct Validity in Psychology. *Philosophy of Science*, *86*(5), 1250–1261. https://doi.org/10.1086/705567

Strauss, V. (2015, March 11). The rise of the anti-standardized testing movement. *Washington Post*. https://www.washingtonpost.com/news/answer-sheet/wp/2014/10/30/the-rise-of-the-anti-standardized-testing-movement/

Terman, L. M. (1916a). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale*. Houghton Mifflin Company.

Terman, L. M. (1916b). The Binet Scale and the Diagnosis of Feeble-Mindedness. *Journal of the American Institute of Criminal Law and Criminology*, *7*(4), 530–543. https://doi.org/10.2307/1133997

Terman, L. M. (1919). *The intelligence of school children: How children differ in ability, the use of mental tests in school grading and the proper education of exceptional children*. Houghton, Mifflin & Company.

Terman, L. M. (1924). The Mental Test as a Psychological Method. *Psychological Review*, *31*(2), 93–117. https://doi.org/10.1037/h0070938

Terman, L. M., & Merrill, M. A. (1937). *Measuring intelligence: A guide to the administration of the new revised Stanford-Binet tests of intelligence.* Houghton Mifflin company.

Thorndike, E. L. (1926). *The measurement of intelligence*. Bureau of Publications, Teacher's College, Columbia University.

Thurstone, L. L. (1924/2013). *The Nature of Intelligence*. Routledge. https://doi.org/10.4324/9781315010298

Thurstone, L. L. (1937). *The reliability and validity of tests; derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems*. Edwards Brothers, Inc.

Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of The vectors of the mind.* The University of Chicago Press.

Tucker, W. H. (2024). *"The Bell Curve" in Perspective: Race, Meritocracy, Inequality and Politics*. Springer Nature Switzerland.

Warne, R. T., Astle, M. C., & Hill, J. C. (2018). What do undergraduates learn about human intelligence? An analysis of introductory psychology textbooks. *Archives of Scientific Psychology*, *6*(1), 32–50. https://doi.org/10.1037/arc0000038

Wechsler, D. (1939). *The measurement of adult intelligence*. The Williams & Wilkins company.

Wechsler, D. (1941). *The measurement of adult intelligence* (2nd ed.). The Williams & Wilkins company.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence.* (4th ed.). Williams & Wilkins.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.