

An Analytical Study on the Use of Artificial Intelligence in Measuring Human Intelligence

Vedanth R. ¹ & Soumya Prakash B. S. ²

¹ Research Intern, Centre for Educational and Social Studies, Bengaluru and Student of STG Pre University College, Mandya, Karnataka, India,
ORCID: 0009-0002-6875-0329; Email: y14223826@gmail.com

² Associate Professor, Centre for Educational and Social Studies, No.6/6, "Prajnanam", 10th Block, II Stage, Nagarabhavi, Bengaluru - 560 072, Karnataka, India,
ORCID: 0009-0008-0744-4401; Email: soumyabs@cessedu.org

Area/Section: Technology Application.

Type of Paper: Exploratory Research.

Number of Peer Reviews: Two.

Type of Review: Peer Reviewed as per [C|O|P|E|](#) guidance.

Indexed in: OpenAIRE.

DOI: <https://doi.org/10.5281/zenodo.22981476>

Google Scholar Citation: [PIJET](#)

How to Cite this Paper:

Vedanth, R. & Soumya Prakash, B. S. (2026). An Analytical Study on the Use of Artificial Intelligence in Measuring Human Intelligence. *Poornaprajna International Journal of Emerging Technologies (PIJET)*, 3(2), 132-140. DOI: <https://doi.org/10.5281/zenodo.22981476>

Poornaprajna International Journal of Emerging Technologies (PIJET)

A Refereed International Journal of Poornaprajna Publication, India.

ISSN: 3107-8486

Crossref DOI: <https://doi.org/10.64818/PIJET.3107.8486.0028>

Received on: 12/07/2026

Published on: 28/09/2026

© With Authors.



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](#), subject to proper citation to the publication source of the work.

Disclaimer: The scholarly papers as reviewed and published by Poornaprajna Publication (P.P.), India, are the views and opinions of their respective authors and are not the views or opinions of the PP. The PP disclaims of any harm or loss caused due to the published content to any party.

An Analytical Study on the Use of Artificial Intelligence in Measuring Human Intelligence

Vedanth R. ¹ & Soumya Prakash B. S. ²

¹ Research Intern, Centre for Educational and Social Studies, Bengaluru and Student of STG Pre University College, Mandya, Karnataka, India,

ORCID: 0009-0002-6875-0329; Email: y14223826@gmail.com

² Associate Professor, Centre for Educational and Social Studies, No.6/6, "Prajnanam", 10th Block, II Stage, Nagarabhavi, Bengaluru - 560 072, Karnataka, India,

ORCID: 0009-0008-0744-4401; Email: soumyabs@cessedu.org

ABSTRACT

Purpose: This study explores how Artificial Intelligence (AI) has emerged as a tool for measuring human intelligence and ascertains whether the AI-based methods in measurement can supplement, enhance or substitute psychometric evaluation in any way. The study will also explore how the fields of psychometrics and AI are converging, highlight the essential AI-based methods in intelligence assessment, point out their strengths and weaknesses, and discuss the issues of fairness, impartiality, validation, and responsibility in relation to them.

Method/Approach: The research employs a qualitative, theoretical, and interpretative methodology that draws on the narrative-thematic assessment of secondary literature sources. The pertinent articles from peer-reviewed articles and preprints along with institutional reports were subjected to thematic content analysis. The literature was categorised into theme groups as follows: classical psychometric principles, AI-based computerised adaptive testing, brain imaging-based prediction, behaviour and multi-modal signal analysis, AI processes used in the human intelligence tests, and issues of fairness and ethics.

Results/Findings: The results indicate that AI has been shown to be the most effective tool for enhancing the efficiency and accuracy of standard intelligence testing methods in terms of computerized adaptive testing and automated scoring. Methods based on neuroimaging provide

predictions about intelligence that are statistically weak but scientifically valid; behavioral methods such as eye-tracking have proven to be promising but may require substantial resources to be applied effectively. Yet, there is no single AI method that can provide full construct validity, standardization and reliability, and fairness as compared to conventional intelligence tests administered by humans. Moreover, these results once again bring to light issues of algorithmic bias, consent, transparency, and instability of responses in generative AI methods.

Originality/Value: This research offers a coherent framework for comparing research studies on different aspects of AI in psychometrics, adaptive testing, neuroimaging, behavioral inference, and evaluation of AI technology. It presents a hybrid assessment model involving AI in adaptive delivery, scoring, and signal extraction while allowing qualified professionals to stay in charge of interpreting results and ensuring ethical integrity.

Type of Paper: Exploratory Research.

Keywords: Artificial Intelligence; Human Intelligence; Psychometrics; Cognitive Assessment; Computerized Adaptive Testing.

1. INTRODUCTION :

1.1 Background:

The scientific measurement of human intelligence began with Alfred Binet and Théodore Simon's 1905 scale, designed to identify French schoolchildren needing academic support (Binet & Simon (1905). [1]). Charles Spearman's proposal of a general intelligence factor "g" gave the enterprise its statistical

backbone (Spearman (1904). [2]), and later scholars extended the theory through Cattell's (1963) [3] fluid-crystallized distinction, Carroll's (1993) [4] three-stratum model, Gardner's (1983) [5] theory of multiple intelligences, and Sternberg's (1985) [6] triarchic theory. David Wechsler's (1939) [7] deviation-IQ scales became the clinical standard, and item-response theory (IRT) later supplied the statistical machinery for reliability, standardization, and norming that still defines a psychometrically defensible test (Embretson & Reise, (2000). [8]). Artificial intelligence has, over the past decade, moved from a peripheral computing curiosity into a central tool across nearly every applied science. Its intersection with intelligence measurement is conceptually distinctive: AI is not merely a new instrument for measuring intelligence, it is itself frequently defined and evaluated with reference to the very psychometric tests built for humans (Bringsjord & Schimanski (2003). [9]). This dual relationship—AI as a measurement tool for human ability, and human tests as a benchmark for AI ability, is the animating tension of this study. Contemporary applications span computerized adaptive testing driven by machine learning (Liu et al. (2024). [10]), neuroimaging-based prediction, and large-language-model evaluation.

1.2 Problem Statement:

Traditional psychometric assessment is costly, time-intensive, dependent on trained examiners, and vulnerable to cultural and linguistic loading, while static item banks are increasingly susceptible to memorization and leakage in a digitally connected world. AI-based alternatives are being adopted rapidly by commercial platforms, educational institutions, and research laboratories, often faster than the scientific validation, fairness auditing, and ethical governance needed to support such adoption (Cheng (2026). [11]; Ganuthula & Balaraman, (2025). [12]). At present, the literature on AI-based intelligence measurement is fragmented across separate sub-disciplines psychometric AI, neuroimaging-based machine learning, affective and behavioural computing, and large-language-model evaluation with no integrative account that maps these strands against one another or against the century-old validity standards of human psychometrics. This study addresses that fragmentation.

1.3 Research Objectives:

- (1) To trace the theoretical convergence of psychometrics and artificial intelligence in the measurement of human intelligence.
- (2) To classify and map the principal AI-based approaches currently used to measure human intelligence.
- (3) To evaluate the comparative accuracy, strengths, and limitations of these approaches against established psychometric standards.
- (4) To examine the fairness, bias, and ethical concerns arising from AI-based intelligence testing.

1.4 Scope and Limitations:

RQ1: What theoretical and technological developments have enabled the application of AI to the measurement of human intelligence?

RQ2: What are the principal AI-driven approaches, adaptive testing, neuroimaging-based prediction, behavioural-signal inference, and LLM-based evaluation, currently used to measure intelligence?

1.5 Scope and Limitations:

The study's scope is confined to peer-reviewed journal articles, preprints, and institutional reports published broadly between 2000 and 2026 that address the application of AI to the measurement of human intelligence across psychometrics, neuroimaging, and behavioural computing. As a conceptual, literature-based study, it does not collect or analyse primary data from human respondents; its conclusions are therefore interpretive syntheses rather than newly generated statistical findings. The pace of AI development means that specific accuracy figures reported here may shift as new model architectures emerge. The review is limited to English-language sources and focuses on AI as a measurement tool for human cognitive ability, referencing the evaluation of AI systems themselves (e.g., large language models) only where it illuminates the underlying measurement question.

1.6 Significance of the Study:

The study is intended to be of value to psychologists and psychometricians assessing whether AI tools meet established validity standards, to educators and human-resource professionals considering AI-based cognitive-assessment products, to policymakers drafting governance frameworks for algorithmic testing, and to AI researchers seeking a psychometrically grounded account of what their models are and are not capturing when they claim to measure intelligence.

2. REVIEW OF LITERATURE :

2.1 Classical Foundations of Intelligence Measurement :

The psychometric tradition begins with Spearman's (1904) [2] discovery that performance across diverse cognitive tasks correlates positively, giving rise to the general intelligence factor, Binet and Simon (1905) [1] operationalized intelligence as a developmental, age-referenced construct, a lineage extended by Wechsler (1939) [7], whose deviation-IQ approach and separate verbal and performance scales became the clinical gold standard for decades. Cattell (1963) [3] divided general ability into fluid intelligence (novel problem-solving) and crystallized intelligence (accumulated knowledge), later integrated by Carroll (1993) [4] into the three-stratum Cattell-Horn-Carroll (CHC) model that still structures most contemporary test batteries. Notably, pluralistic accounts, such as Gardner's (1983) [5] multiple intelligences and Sternberg's (1985) [6] triarchic theory, challenged the primacy of a single g factor, broadening the construct to include practical, creative, and domain-specific abilities. Embretson and Reise (2000) [8] subsequently formalized item-response theory (IRT), supplying the statistical foundation of reliability, standardization, and item calibration on which nearly all AI-based adaptive-testing systems now depend.

2.2 Psychometric AI: The Conceptual Bridge:

Bringsjord and Schimanski (2003) [9] argued for a striking conceptual claim: that the very definition of artificial intelligence is best understood through psychometrics, proposing "Psychometric AI" as an entity capable of solid performance on the full range of validated human ability tests, not only IQ tests but also tests of creativity and mechanical aptitude. This reframing collapses the usual distinction between AI as a tool for measuring human intelligence and AI as a system whose own intelligence is measured by the same instruments, a duality that recurs throughout the more recent literature reviewed below.

2.3 AI-Driven Computerized Adaptive Testing (CAT):

Computerized adaptive testing, which selects each subsequent item based on prior responses using IRT, is the most mature and widely deployed AI application in this domain. Liu et al. (2024) [10] provide a machine-learning-oriented survey of CAT, examining measurement models, question-selection algorithms, item-bank construction, and cognitive diagnosis models, and note that CAT has expanded from education into healthcare, sports, and the evaluation of AI systems themselves. Adaptive engines concentrate test items near an examinee's true ability level, which compared with fixed-form tests reduces test length substantially while maintaining or improving measurement precision, particularly at the extremes of the ability distribution where fixed-form tests are traditionally least informative. Applied examples include adaptive inductive-reasoning batteries built on two-parameter logistic IRT models and consumer-facing adaptive cognitive assessments that reference established frameworks such as the Wechsler scales and CHC theory for score calibration.

2.4 Neuroimaging-Based AI Prediction of Intelligence:

A parallel body of work asks whether intelligence can be predicted directly from brain structure using deep learning. Choi et al. (2008) [13] linked cortical thickness and neural activation patterns to multiple bases of human intelligence, providing an early neuroscientific anchor for later computational work. The Adolescent Brain Cognitive Development Neurocognitive Prediction Challenge became a focal benchmark: across more than twenty submitted models, prediction of residualized fluid intelligence from structural MRI produced mean-square errors in the range of roughly 86 to 103, indicating that image-derived features carry only modest predictive signal for individual differences in ability. Saha, Pagnozzi, Bradford, and Fripp (2021) [14] applied convolutional neural networks to structural MRI and established a correlation between brain features and raw fluid intelligence that was statistically stable but numerically weak. Later architectures including three-dimensional U-Nets enhanced with residual

and squeeze-and- excitation blocks, and graph convolutional networks operating on cortical and subcortical surface meshes improved on these baselines, with combined cortical-subcortical models reaching correlations of roughly $r = 0.45$ for fluid-intelligence prediction, and related work extending this approach to verbal, performance, and full-scale IQ using structural imaging. Across this literature, the consistent conclusion is that brain-structure-based AI prediction is scientifically informative about the neural correlates of ability but remains far too imprecise for individual-level diagnostic use.

2.5 Behavioural and Multimodal AI Measurement:

A third strand moves away from brain imaging and test items altogether, inferring cognitive-affective constructs from unobtrusive behavioural signals. Wang et al. (2023) [15] trained seven machine-learning classifiers on eye-tracking data to measure facets of emotional intelligence, running roughly 1,470 model evaluations and achieving average accuracies between 50% and 90% depending on the scoring threshold used; notably, high accuracy was attainable from as little as two to five seconds of gaze data. This line of research suggests that AI can extract psychometrically meaningful signal from passive, low-friction behavioural streams rather than requiring an examinee to complete a lengthy explicit test, though it also raises new questions about consent and the boundaries of what counts as a "test" at all.

2.6 AI Systems as Subjects of Human Intelligence Tests:

A distinct and rapidly growing literature inverts the usual direction of measurement, administering human-normed intelligence tests to AI systems. Klein and Kovacs (2024) [16] gave a computerized adaptive vocabulary test to ChatGPT (GPT-3.5) and Bing (GPT-4), finding that both systems performed at a level exceeding roughly 95% of human test-takers, yet gave different answers to the same repeated item in 42% of cases and, when wrong, tended to hallucinate answers outside the given options rather than guess a response pattern that classical psychometric theory, built around stable human traits, did not anticipate. Wang, Jiang, Hernandez-Orallo, Stillwell, Sun, Luo, and Xie (2023) [17] similarly proposed evaluating general-purpose AI systems using psychometric methodology, arguing that established ability tests offer a rigorous complement to conventional machine-learning benchmarks. Together, this literature suggests that the validity of applying human-calibrated instruments to non-human systems and, reciprocally, of trusting AI-graded interpretations of human performance, cannot be assumed and requires independent verification.

2.7 Applied and Commercial AI Intelligence-Measurement Tools:

Beyond the academic literature, applied frameworks have begun reframing intelligence measurement for an AI-saturated environment. Ganuthula and Balaraman (2025) [12] proposed an "Artificial Intelligence Quotient" (AIQ) framework, arguing that classical human intelligence tests need to be reconceived to capture a person's ability to collaborate effectively with AI systems rather than solely their unaided cognitive performance; the authors cite survey evidence that a large share of the public cannot accurately describe common AI applications, underscoring a literacy gap that existing intelligence batteries were never designed to capture. Consumer-facing adaptive cognitive-assessment products have proliferated in parallel, typically combining adaptive item delivery with automated scoring referenced against established frameworks such as the Wechsler Adult Intelligence Scale and CHC theory, though such tools generally have far less independent psychometric validation than the academic instruments they draw upon.

2.8 Fairness, Bias, and Ethics in AI-Based Cognitive Assessment:

As AI-based assessment enters higher-stakes domains such as hiring and education, questions of fairness have become central. Fairness has become central. Cheng (2026) [11] offers a systematic mapping of the longstanding test-fairness tradition in psychometrics onto the algorithmic-fairness paradigm in machine learning, arguing that fairness must be evaluated across the entire testing workflow, including item writing, scoring, and use, rather than only at a final selection decision. Building on this, Xu and Ong (2025) [18] propose a "Fair-IRT" framework that applies item-response theory to evaluate fairness in machine-learning models directly, while the wider algorithmic-fairness literature distinguishes group fairness (equal outcomes across demographic groups) from individual fairness (similar treatment of similar individuals), and separates data bias, arising from

unrepresentative training samples, from algorithmic bias, arising from the model-fitting process itself. This body of work collectively signals that fairness auditing for AI-based cognitive assessment is an active but still-maturing area, with much of the foundational framework literature dated 2024 to 2026.

2.9 Research Gap:

Reviewing this literature side by side surfaces four interlinked gaps that the present study addresses. First, the literature is siloed: psychometric-AI theory, adaptive-testing engineering, neuroimaging-based prediction, behavioural-signal inference, and LLM evaluation are rarely discussed within a single integrative frame, making it difficult for a practitioner or policymaker to compare their relative maturity and accuracy. Second, there is no settled account of construct validity across methods, it remains unclear whether adaptive-test scores, brain-imaging predictions, and gaze-based inferences are measuring the same underlying construct of "intelligence" or distinct, only loosely related proxies of it. Third, the empirical accuracy of neuroimaging-based prediction remains modest (correlations generally below 0.5), yet this limitation is not always transparently communicated in applied or commercial contexts. Fourth, fairness and bias frameworks specific to AI-based cognitive assessment are only now emerging (predominantly 2024-2026), lagging well behind the pace of commercial deployment of AI-based testing products. This study responds to these gaps by offering an integrative, comparative synthesis and a proposed hybrid framework for responsible use.

3. RESEARCH METHODOLOGY :

3.1 Research Design:

This study adopts a qualitative, conceptual, and interpretive research design, structured as a narrative-thematic literature review. Given the heterogeneity of the underlying literature spanning quantitative neuroimaging studies, machine-learning benchmark papers, and conceptual/ theoretical psychometric articles, a strict quantitative meta-analysis was judged inappropriate; instead, the design follows a systematic identification screening, synthesis and logic adapted from standard review methodology, oriented toward theory-building rather than effect-size pooling.

3.2 Source of Data:

All data used in this study are secondary in nature. Material was drawn from peer-reviewed journal articles and institutional reports, accessed through open-access databases such as PubMed/PMC, ScienceDirect, SpringerLink, and Google Scholar Sources were selected for direct relevance to the intersection of artificial intelligence and human-intelligence measurement, spanning publication years from the foundational psychometric literature of the early twentieth century through contemporary AI research up to 2026.

3.3 Method of Analysis:

The gathered literature was analyzed using thematic content analysis. Sources were first coded into six emergent thematic categories: classical psychometric foundations; the psychometric-AI conceptual bridge; AI-driven computerized adaptive testing; neuroimaging-based prediction; behavioural and multimodal signal-based inference; and AI systems as test subjects together with associated fairness and ethical concerns. Within each theme, reported accuracy, validity, and reliability indicators were extracted and compared to address the study's research questions, after which an inductive synthesis step was used to identify the cross-cutting research gap and to construct the integrated framework and recommendations presented in the sections that follow.

4. MAIN ANALYSIS / DISCUSSION OF THE STUDY :

4.1 From Static to Adaptive:

AI's Clearest Efficiency Gain. Across the reviewed literature, AI-driven computerized adaptive testing represents the most mature and best-validated application of AI to intelligence measurement. By selecting items dynamically around an examinee's estimated ability using item-response theory, CAT systems achieve large reductions in test length while preserving or improving measurement precision, particularly for individuals at the high or low ends of the ability distribution who are poorly served by fixed-form tests (Liu et al., 2024) [10]. This is best understood not as AI replacing psychometric theory, but as AI operationalizing it more efficiently.

4.2 The Brain as Data:

Promise and Limits of the Neuroimaging Frontier Neuroimaging-based prediction represents the most technically ambitious but empirically humbling AI application reviewed in this study. Despite increasingly sophisticated architectures—three-dimensional convolutional networks, residual and attention mechanisms, and graph-based models operating on cortical surface meshes reported correlations between predicted and actual fluid-intelligence scores remain modest, typically in the range of $r = 0.1$ to 0.45 . This pattern suggests that structural brain measurements alone capture only a fraction of the variance in individual cognitive ability, consistent with theoretical accounts such as the Parieto-Frontal Integration Theory, which frame intelligence as an emergent property of distributed neural interaction rather than a property localisable to any single structural measurement. The practical implication is that, at present, neuroimaging-based AI prediction is a research tool for understanding population-level neural correlates of ability, not a viable individual-level diagnostic instrument.

4.3 Beyond the Test Booklet: Behavioural and Multimodal Inference:

The behavioural-signal literature, exemplified by eye-tracking-based prediction of emotional intelligence, demonstrates that AI can extract psychometrically useful signal from passive data streams collected in seconds rather than the tens of minutes required by conventional testing (Wang et al. (2023) [15]. This points toward a future of low-friction, potentially continuous assessment embedded in everyday digital interaction. At the same time, the very feature that makes this approach attractive—its unobtrusiveness—is also its central ethical liability, since ability inference from passive behavioural monitoring raises materially different consent and transparency questions than a person voluntarily sitting for a conventional test.

4.4 Turning the Test on the Machine:

What AI-as-Examinee Reveals The finding that large language models outperform the great majority of human test-takers on a verbal computerized adaptive test, while simultaneously giving inconsistent answers to identical repeated items and hallucinating options that do not exist (Klein & Kovacs, 2024) [16], is analytically significant beyond its novelty. It demonstrates that psychometric instruments encode assumptions, chiefly, that the examinee's underlying trait is stable across administrations, that hold for humans but do not straightforwardly hold for generative AI systems. This matters directly for the present study's subject: if AI systems administering or scoring human intelligence tests inherit the same architectural properties that produce inconsistency when they are the examinee, the reliability of AI mediated human assessment cannot simply be assumed and must be independently verified for each application.

4.5 Fairness as Unfinished Business:

The fairness and bias literature reviewed in this study (Cheng (2026). [11]; Xu & Ong (2025) [18] shows a field actively importing decades of test-fairness theory from psychometrics into the machine-learning fairness paradigm, but doing so only very recently relative to the pace at which AI-based cognitive-assessment products have already reached commercial markets. The distinction between data bias (unrepresentative training samples) and algorithmic bias (bias introduced during model fitting) is analytically useful but not yet routinely operationalised in applied testing products, and the difference between group fairness and individual fairness has direct consequences for how an AI-based intelligence-assessment tool would need to be audited before deployment in consequential settings such as hiring or special-education placement.

4.6 Synthesis:

Proxy, Not Replacement. Drawing the six themes together, this study's central analytical conclusion is that current AI applications largely measure correlates or proxies of intelligence, adaptive response patterns, structural brain features, gaze dynamics, rather than directly instantiating a unified construct equivalent to Spearman's g [2]. This is not a criticism unique to AI; psychometrics has debated the unity of intelligence since 1904. What is distinctive is that AI methods now allow many different proxies to be measured simultaneously and cheaply, which increases the urgency of construct-validity research to

determine how, and whether, these proxies should be combined, weighted, or kept analytically separate.

5. FINDINGS / RESULTS :

- (1) AI most reliably improves the efficiency and precision of already-validated psychometric methods through computerized adaptive testing, rather than replacing the underlying test theory.
- (2) Neuroimaging-based AI prediction of intelligence is scientifically informative but statistically modest, with correlations generally below $r = 0.5$, making it currently unsuitable for individual-level diagnostic use.
- (3) Behavioural and multimodal AI approaches, such as eye-tracking-based inference, can achieve moderate-to-high accuracy (roughly 50-90%) from only a few seconds of data, indicating strong potential for low-friction continuous assessment.
- (4) Large language models score highly on human-normed verbal intelligence tests yet display response instability across repeated administrations, exposing a mismatch between classical psychometric assumptions and generative-AI behaviour.
- (5) Fairness and bias frameworks tailored to AI-based cognitive assessment are still emergent, with the bulk of the relevant literature dated 2024-2026, trailing behind commercial adoption.
- (6) No single AI-based approach yet reproduces the full construct validity, standardization, and multidimensionality of established human-administered batteries such as the Wechsler scales.
- (7) The evidence base most strongly supports a hybrid human-AI model, in which AI manages adaptive delivery, scoring efficiency, and signal extraction, while trained professionals retain interpretive authority and ethical oversight.

6. RECOMMENDATIONS :

- (1) Deploy AI primarily as an efficiency and augmentation layer atop psychometrically validated instruments, for adaptive delivery and automated scoring, rather than as an independent replacement for established tests.
- (2) Require fairness audits covering both group and individual fairness, following frameworks such as those proposed by Cheng (2026) [11] and Xu and Ong (2025) [18], before any AI-based cognitive-assessment tool is used in high-stakes settings such as hiring, admissions, or clinical evaluation.
- (3) Mandate transparent reporting of accuracy and error metrics for all AI-based cognitive-assessment tools, so that users and stakeholders can evaluate the reliability of results before relying on them in consequential decisions.

7. CONCLUSION :

This study set out to map, compare, and critically evaluate the expanding use of artificial intelligence in measuring human intelligence. The evidence reviewed shows that AI has meaningfully reshaped the practice of intelligence testing, enabling adaptive, efficient delivery through computerized adaptive testing; opening a genuinely new evidentiary channel through neuroimaging-based prediction; and extracting psychometrically relevant signal from previously unusable behavioural data such as eye movement. At the same time, the century-old validity bar set by classical psychometrics—reliability, standardization, cross-population fairness, and demonstrated construct validity has not yet been fully met by any single AI-based approach reviewed here. Neuroimaging predictions remain statistically modest, behavioural inference raises unresolved consent questions, and even the most linguistically capable AI systems display response patterns that classical test theory did not anticipate. The most defensible path forward, and the one this study recommends, is not a race to replace human-administered testing with AI, but the deliberate construction of hybrid systems in which AI contributes efficiency, scale, and new forms of signal, while human expertise continues to anchor interpretation, fairness, and ethical accountability. Measuring the mind, whether by pencil, brain scan, or algorithm, remains as much a matter of scientific humility as of technical capability.

REFERENCES :

- [1] Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, 11(1), 191-244. [Google Scholar](#)[↗]

- [2] Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15(2), 201-292. [Google Scholar↗](#)
- [3] Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1-22. DOI: <https://doi.org/10.1037/h0046743>
- [4] Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge University Press. [Google Scholar↗](#)
- [5] Gardner, H. (1983). Frames of mind: The theory of multiple intelligences. Basic Books. [Google Scholar↗](#)
- [6] Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of human intelligence. Cambridge University Press. [Google Scholar↗](#)
- [7] Wechsler, D. (1939). The measurement of adult intelligence. Williams & Wilkins. [Google Scholar↗](#)
- [8] Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates. [Google Scholar↗](#)
- [9] Bringsjord, S., & Schimanski, B. (2003). What is artificial intelligence? Psychometric AI as an answer. In Proceedings of the 18th International Joint Conference on Artificial Intelligence (pp. 887-893). [Google Scholar↗](#)
- [10] Liu, Q., Zhuang, Y., Bi, H., Huang, Z., Huang, W., Li, J., Yu, J., Liu, Z., Hu, Z., Hong, Y., Pardos, Z. A., Ma, H., Zhu, M., Wang, S., & Chen, E. (2024). Survey of computerized adaptive testing: A machine learning perspective. arXiv preprint arXiv:2404.00712. DOI: <https://doi.org/10.48550/arXiv.2404.00712>
- [11] Cheng, Y. (2026). Fairness issues and evaluation in psychometrics and AI/ML: What can we learn from each field? *Psychometrika*. Advance online publication. DOI: <https://doi.org/10.1017/psy.2026.10110>
- [12] Ganuthula, V. R. R., & Balaraman, K. K. (2025). Artificial intelligence quotient framework for measuring human collaboration with artificial intelligence. *Discover Artificial Intelligence*, 5, Article 268. DOI: <https://doi.org/10.1007/s44163-025-00516-1>
- [13] Choi, Y. Y., Shamos, N. A., Cho, S. H., DeYoung, C. G., Lee, M. J., Lee, J.-M., Kim, S., Cho, Z.-H., Kim, K., Gray, J. R., & Lee, K. H. (2008). Multiple bases of human intelligence revealed by cortical thickness and neural activation. *Journal of Neuroscience*, 28(41), 10323-10329. DOI: <https://doi.org/10.1523/JNEUROSCI.3259-08.2008>
- [14] Saha, S., Pagnozzi, A., Bradford, D., & Fripp, J. (2021). Predicting fluid intelligence in adolescence from structural MRI with deep learning methods. *Intelligence*, 88, Article 101568. DOI: <https://doi.org/10.1016/j.intell.2021.101568>
- [15] Wang, W., Kofler, L., Lindgren, C., Lobel, M., Murphy, A., Tong, Q., & Pickering, K. (2023). AI for psychometrics: Validating machine learning models in measuring emotional intelligence with eye-tracking techniques. *Journal of Intelligence*, 11(9), Article 170. DOI: <https://doi.org/10.3390/jintelligence11090170>
- [16] Klein, B., & Kovacs, K. (2024). The performance of ChatGPT and Bing on a computerized adaptive test of verbal intelligence. *PLOS ONE*, 19(7), Article e0307097. DOI: <https://doi.org/10.1371/journal.pone.0307097>
- [17] Wang, X., Jiang, L., Hernandez-Orallo, J., Stillwell, D., Sun, L., Luo, F., & Xie, X. (2023). Evaluating general-purpose AI with psychometrics. arXiv preprint arXiv:2310.16379. DOI: <https://doi.org/10.48550/arXiv.2310.16379>
- [18] Xu, Z., & Ong, C. S. (2025). Fairness evaluation with item response theory. In Proceedings of the ACM Web Conference 2025 (pp. 2276-2288). DOI: <https://doi.org/10.1145/3696410.3714883>
