MAIMON WORKING PAPER No 11 AUGUST 2025

REJECTING NUMERICAL STORYTELLING: APPLYING RASCH MEASUREMENT STANDARDS FOR A NEW AUSTRALIAN IQ PROFILE INSTRUMENT

Paul C. Langley Ph.D. Adjunct Professor, College of Pharmacy, University of Minnesota, Minneapolis, MN

ABSTRACT

In Australia, the Wechsler Adult Intelligence Scale (WAIS) remains enshrined as the benchmark for assessing cognitive ability across clinical, educational, and legal arenas. Yet its foundational use of ordinal data, norm-referenced scoring, and profile interpretation renders it epistemologically invalid. WAIS scores, including subtests, index scores, and the Full Scale IQ result from ordinal rankings normed against age groups. These scores violate core axioms of scientific measurement: unidimensionality, invariance, and linear, interval-level scaling. They do not measure latent intelligence; rather, they narrate rank positions, a form of numerical storytelling masquerading as quantification.

The upcoming WAIS-5 A&NZ edition, despite offering updated norms and revised indices, perpetuates these conceptual flaws. Without true zero points or guarantee of equal intervals, the WAIS-5 cannot produce meaningful arithmetic operations essential for scientific inference. Therefore, claims like the Flynn Effect, assuming generational IQ shifts, stand on unstable ground, given they derive from instruments lacking interval or ratio scale properties.

Rasch measurement theory offers the only plausible pathway to valid latent trait measurement. Unlike WAIS's arbitrary scoring conventions, Rasch modeling operationalizes a probabilistic, invariant structure that maps person ability and item difficulty onto a common logit scale. This transformation yields additive and linear measures, but only if three conditions are met: unidimensional latent constructs, sample- and item-invariant functioning, and rigorous fit to the model's expectations

In the Australian context, the improper use of WAIS profile scores carries significant ethical ramifications. Access to NDIS services, eligibility for educational support, and legal determinations often hinge on these pseudo-measures. Reliance on non-scientific scoring mitigates validity, misclassifies individuals, and undermines professional integrity.

This paper critiques WAIS's structural failures and argues for a transition to Rasch-based instruments tailored to Australian settings. Rather than generating a single Full Scale IQ derived from aggregated, non-equivalent subtests, a Rasch instrument would deliver unidimensional, trait-specific measures grounded in scientific rigor. Built through careful item development, calibration, and validation across populations, such tools would support meaningful, interpretable measurement of cognitive constructs.

Implementing this shift necessitates institution-wide commitment, from professional education reforms to endorsement by regulatory bodies like APS and AHPRA. Training programs must foreground measurement theory; regulators should require evidence of Rasch conformity; and clinical practice must pivot from convenience-driven testing toward scientific measurement.

Without such transformation, Australian psychology risks remaining mired in narrative, not knowledge. Rasch measurement offers a genuine scientific strategy, one that restores integrity, yields meaningful trait quantification, and aligns psychological assessment with the standards of real measurement.

INTRODUCTION

The practice of psychological assessment in Australia continues to rely on outdated and epistemologically flawed instruments to evaluate cognitive ability. Chief among these is the Wechsler Adult Intelligence Scale (WAIS), a tool deeply embedded in clinical, educational, and legal contexts [1]. Despite its widespread use globally and specifically in Australia, the WAIS and its associated interpretive practice of profile analysis that fails to meet the basic requirements of scientific measurement [2] [3] [4]. Its subtests and index scores are constructed from ordinal data, norm-referenced against age cohorts, and aggregated in ways that violate the axioms of unidimensionality, invariance, and scale linearity. The result is not a measure of intelligence, but a descriptive scoring system misrepresented as quantification. This paper argues that continued reliance on WAIS profile scores constitutes a form of numerical storytelling, where the appearance of measurement substitutes for its substance and proposes an alternative grounded in Rasch measurement theory [5] .

The *Wechsler Adult Intelligence Scale, Fifth Edition: Australian and New Zealand Language Standardised Edition* (WAIS-5 A&NZ) is scheduled for release by Pearson Clinical Australia in mid to late 2025. This edition introduces updated local norms, additional subtests, and revised index scores intended to enhance clinical interpretability and efficiency. Yet despite these refinements, the WAIS-5 remains anchored in an ordinal measurement framework. As with previous editions, it fails to satisfy the axioms of fundamental measurement, specifically, the requirements of equal intervals and a true zero. Without these properties, score differences lack ratio meaning and cannot support the arithmetic operations foundational to scientific inference. This renders the WAIS-5 incapable of producing interval or ratio scales, and therefore incapable of yielding quantitative measures of intelligence in any scientific sense.

Claims regarding generational trends in IQ, such as the so-called Flynn Effect, are undermined by these foundational deficiencies [6]. An effect inferred from ordinal scores cannot be assumed to reflect changes in a true underlying quantity, as the WAIS-5 does not measure such a quantity. Indeed, any attempt to interpret an upward or downward shift in mean IQ scores is speculative at best, and scientifically incoherent at worst, when derived from an instrument that has never established interval properties. The persistence of the Flynn Effect as a narrative device in psychometrics is not evidence of latent intelligence growth, but rather a symptom of decades-long indifference to the measurement limitations of the WAIS framework.

Rasch measurement provides the only known pathway to transforming ordinal responses into linear, invariant measures of latent trait possession [7]. Unlike legacy instruments, which derive scaled scores through arbitrary statistical conventions, Rasch modeling constructs a mathematical structure in which person ability and item difficulty are located on a common logit scale [8]. This structure allows for meaningful statements about the degree to which an individual possesses a given trait. It also requires that each instrument satisfy strict criteria: unidimensionality of the latent construct, invariance of item functioning across populations, and the ability to yield additive, linear measures. These are not optional enhancements but necessary conditions for scientific measurement. These requirements, reflecting the agreed axioms of fundamental measurement have been known for over 50 years with the representational theory of measurement [9] .

In Australia, the implications of failed measurement are far-reaching. WAIS scores are used to determine access to public services, support eligibility under the National Disability Insurance Scheme (NDIS), guide educational placement, and provide evidence in legal proceedings. The misuse of profile scores in these contexts risks misdiagnosis, misclassification, and misallocation of resources. More critically, it places psychological authority on a foundation that cannot withstand epistemic scrutiny. The failure is not just technical; it is ethical. Psychologists who base professional judgments on non-measures compromise the validity of their claims and undermine the integrity of the discipline.

This paper sets out a path forward. It begins by critiquing the structure and use of the WAIS in Australian settings, demonstrating its inability to support claims of latent trait measurement. It then introduces the Rasch model as a scientifically defensible alternative and outlines the essential steps required to develop a new, Rasch-based cognitive profile instrument tailored to Australian populations. Such an instrument would abandon the reductive fiction of the full-scale IQ and instead offer a set of unidimensional, trait-specific measures that meet the demands of both clinical utility and scientific rigor. This transformation will require institutional commitment, from training programs, professional bodies, and regulators alike, but it is the only path to reclaiming measurement as a legitimate scientific endeavor in Australian psychology. The stakes are high: without measurement, psychology cannot claim to know; it can only narrate. Rasch offers a way to know.

AUSTRALIAN APPLICATIONS AND FAILURES OF WAIS

In Australia, the Wechsler Adult Intelligence Scale (WAIS) remains the dominant instrument for assessing intelligence in clinical, educational, vocational, and legal settings. Despite its widespread use and procedural standardization, the WAIS fundamentally fails to meet the requirements of scientific measurement. Its outputs, Full Scale IQ (FSIQ) and associated index scores, are norm-referenced, ordinal ranks that reflect population standing rather than any measurable quantity of latent trait possession. Yet these scores are routinely interpreted as if they offered meaningful, linear assessments of cognitive ability.

Psychologists trained under the Australian Psychological Society (APS) and regulated by AHPRA are taught to interpret beyond the FSIQ, drawing on the Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed indices. Discrepancies among these are treated as evidence of specific strengths or deficits. In educational and vocational contexts, such

profiles are used to determine eligibility for services. In legal settings, IQ thresholds may influence rulings on competency or culpability. And in clinical environments, WAIS profiles inform diagnoses of intellectual disability, brain injury, and dementia.

Yet all these applications depend on the assumption that WAIS scores measure something real, which they do not. The FSIQ is constructed from raw scores converted to scaled scores via age-based norms, then aggregated. This process lacks any demonstration of unidimensionality, invariance, or interval scaling. The subtests are not validated as measuring a common construct; index scores are composites of non-equivalent parts; and profile-based interpretations rest on score discrepancies that may be psychometrically meaningless. No part of the WAIS conforms to the axioms of fundamental measurement.

The continued use of the WAIS is not a mark of scientific credibility but a monument to disciplinary inertia. Its foundation lies in the operationalist definition of measurement introduced by Stevens, who in 1946 defined measurement as "the assignment of numerals to objects or events according to rule" [10] While Stevens classified ordinal scaling as legitimate measurement, he was explicit that only non-parametric statistics, medians, percentiles, rank-order coefficients, were appropriate for such data. He rejected the use of means, standard deviations, and correlation coefficients for ordinal scores. Yet in a stunning act of methodological opportunism, mainstream psychology embraced the definition while discarding the constraints. Instruments like the WAIS, which yield rank-ordered scores without demonstrating equal intervals or a true zero, became statistical workhorses, subjected to analyses Stevens himself warned against. This abandonment of even operationalist integrity coincided with the neglect of more rigorous models advanced by Rasch with Luce and Tukey, who insisted that meaningful measurement required unidimensionality, additivity, and invariance [11]. These models laid the groundwork for interval or ratio measurement through structural modeling, not arbitrary number assignment; the representational theory of measurement for physical features as well as latent traits. The WAIS ignores all of this. Its persistence owes nothing to science and everything to professional convenience and the rhetorical misuse of Stevens' name to justify mathematically indefensible practices.

These developments, which offered psychology a path toward scientific legitimacy, were bypassed in favor of normative scoring systems easily integrated into clinical workflows and commercial products. The WAIS has thus endured as a pseudo-measure: statistically sophisticated, institutionally accepted, but epistemologically hollow. The IQ score produced does not quantify intelligence; it positions an individual within a reference group. It is a label of relative standing, not a measure of trait possession.

While the APS promotes profile-based interpretation as a more nuanced approach, this too collapses under scrutiny. The assumption that index differences represent meaningful cognitive disparities is unjustified when those indices are not anchored in a common measurement framework. What appears as diagnostic insight is often the artifact of flawed scale construction. In real-world practice, these errors are not benign. WAIS scores inform judgments in domains, legal, clinical, educational, where validity should be non-negotiable.

The continued use of the WAIS in Australia, therefore, illustrates a deeper failure within psychological science: a refusal to ground its instruments in the logic of fundamental measurement. The necessary tools have existed for over seventy years. Rasch measurement provides a viable alternative, allowing for the construction of trait-based cognitive instruments that are linear, invariant, and interpretable. The failure to transition is not technical, but cultural. Until that failure is addressed, WAIS will remain a sophisticated diagnostic convention masquerading as science.

UNDERSTANDING RASCH TRAIT POSSESSION

There are only two types of scientifically defensible measures. The first is the manifest linear ratio measure, which quantifies observable events or quantities using units with constant absolute differences—such as seconds, grams, or number of hospital admissions. These measures, grounded in direct observation, allow for arithmetic operations and invariant comparisons. The second, and more complex, is the Rasch logit ratio measure, which quantifies possession of a latent construct through a transformation of ordinal response data into a linear scale of constant relative differences. It is only through these two forms, manifest linear ratios and Rasch logits, that measurement, in the scientific sense, is possible. Any other numerical expression, however complex or statistically derived, fails to meet the axioms of fundamental measurement.

In psychological and educational assessment, we are rarely dealing with directly observable quantities. Instead, we infer underlying attributes, such as verbal reasoning, mathematical aptitude, anxiety, or working memory, through observable behaviors captured in item responses. These are manifestations of latent constructs, attributes that cannot be directly observed but which are assumed to explain the variability in response behavior across individuals. The challenge is to construct a measurement framework that allows these unobservable attributes to be meaningfully and consistently quantified.

The Rasch model provides this framework. Unlike traditional test theory, which treats raw scores or composite indices as stand-ins for trait possession, the Rasch model requires that we explicitly define the latent construct, design items to reflect increasing levels of difficulty or endorsement, and then calibrate both persons and items on the same underlying scale. What results is not a score in the ordinal sense, but a logit, the natural logarithm of the odds of a successful response, conditioned on the interaction between person ability and item difficulty; the logit is defined as $\ln(P/1-P)$ where ln is the natural logarithm or base e and P is probability. .

The instrument that supports Rasch measurement begins with item development. For any latent construct, say, verbal abstraction, we must create a series of items that each reflect a different level of the attribute. These items must be unidimensional, meaning they all tap into the same latent variable without contamination from unrelated traits. Once developed, these items are administered to a sample of respondents. Their responses, typically recorded as correct/incorrect or agreement levels, are ordinal by nature. A person who answers five items correctly is ranked higher than one who answers three, but the difference between them is not necessarily meaningful or consistent across the scale. Raw scores do not tell us how much more of the trait one person possesses than another; they tell us only who performed better.

Rasch analysis transforms this raw response matrix into a measurement model by estimating two sets of parameters: person ability and item difficulty. The probability of a correct response (or higher endorsement) is modeled as a function of the difference between the person's ability and the item's difficulty. This relationship is expressed in log-odds units, the logit, which allows the response probabilities to be placed on an additive, linear scale. A person whose ability matches the item difficulty will have a 50% chance of success; a person with higher ability than the item's difficulty will have a higher probability, and vice versa. The beauty of this structure is its invariance: item difficulties do not depend on the sample, and person abilities do not depend on the particular set of items, provided they fit the required Rasch model standards.

This transformation is not cosmetic. It represents a fundamental shift in what it means to measure a trait. Rather than summing item scores and comparing them to normative tables, we now have a calibrated scale where we can make precise, replicable statements about the degree of trait possession. The scale has a true zero, defined probabilistically, and equal units across its continuum. Unlike norm-referenced scores, which are meaningful only relative to the reference group, Rasch measures are meaningful in themselves. A logit difference of 1.0 represents the same magnitude of difference anywhere on the scale.

Moreover, Rasch measurement allows for sophisticated diagnostics. We can test for item fit, ensuring that each item conforms to the expected response pattern. We can assess whether items are functioning equivalently across subgroups, a property known as differential item functioning (DIF). We can examine whether the scale is truly unidimensional, using residual analyses to detect underlying dimensions. All of this serves to protect the validity of the measure and to ensure that what we are quantifying is indeed a single latent trait and not a composite or a statistical fiction.

The implications are profound. Once a Rasch scale has been constructed and validated, we can track individual trait possession over time, compare individuals meaningfully across settings, and evaluate the effects of interventions. A change in logit value is interpretable as a change in trait possession, not merely a change in test performance. This is particularly important in clinical and educational settings, where judgments about therapy effectiveness, cognitive change, or functional capacity must be based on something more than arbitrary score differences.

The key to assessing latent trait status and the responses to interventions is the possession measure. This determines, from the pattern of item responses, the extent to which a respondent or a target group average (and distributional) possess of the latent trait in question (e.g., mathematical ability). It is an interval measure that supports statistic operations and can be transformed to a ratio measure for more intensive evaluations.

To speak of Rasch trait possession is to speak of a legitimate scientific claim. It allows us to quantify the unobservable in a way that supports inference, comparison, and hypothesis testing. It restores the integrity of measurement in domains long plagued by numerical storytelling. No other approach offers this combination of conceptual rigor and practical utility. Classical test theory is exploratory and descriptive, but not measurement. Factor analysis offers dimensional reduction, but not linear scale construction. Only Rasch offers measurement in the same sense that physics or chemistry understands the term: a structured, invariant, mathematically grounded representation of quantity; the Rasch model is confirmatory and predictive.

In replacing traditional IQ tests and their ordinal derivatives with Rasch-based instruments, we make a decisive epistemic shift. We move from normative illusions to structural representations. We cease ranking people against each other and begin measuring them along the continuum of trait possession. This is not merely a change in technique. It is a transformation in what it means to know something scientifically about human attributes. In the Australian context, where psychological assessments are often entangled with access to services, education, and legal rights, this shift is not optional. It is a moral and scientific imperative.

For those who fear the challenge of Rasch modelling there are a range of statistical packages (e.g., WINSTEPS) that support Rasch. These have been available for some 35 years. They provide the basis for instrument development, the fit of items to the Rasch model and an evaluation of possession in logits.

THE MISUSE OF PROFILE SCORES

The widespread use of profile-based interpretation in Australian psychological practice, particularly with instruments such as the WAIS-4, reflects a systemic failure to engage with the principles of fundamental measurement. Although profile analysis is framed as a refined, individualized approach to cognitive assessment, it is built on a conceptual void. The practice assumes that scores derived from WAIS subtests and indices represent measurable attributes of cognitive traits, yet the numerical outputs are not, and cannot be, defended as measures in any scientific sense. The entire structure rests on the unquestioned legitimacy of summing ordinal scores, converting them to norm-referenced scales, and then drawing interpretive conclusions based on presumed quantitative differences. This is not measurement; it is inference without epistemic justification.

Profile interpretation typically involves comparing index scores (e.g., Verbal Comprehension Index versus Processing Speed Index), identifying discrepancies between subtest results, and making diagnostic suggestions based on presumed cognitive strengths and weaknesses. Psychologists trained under the guidance of the Australian Psychological Society (APS) and regulated by the Australian Health Practitioner Regulation Agency (AHPRA) are encouraged to go beyond reporting the FSIQ and engage in this deeper analysis of subtest scatter. It is presented as a clinical strength, offering a nuanced view of the individual's cognitive profile. Yet, behind this practice lies a profound misconception: the belief that the numbers produced by the WAIS-4 have ratio or even interval properties that support the arithmetic operations required for valid interpretation.

In fundamental measurement theory, one cannot legitimately claim that a higher score means "more" of an attribute unless that score is placed on a linear scale with either constant absolute or constant relative differences. This foundational requirement is ignored in WAIS-derived profile analysis. The raw scores, often summed over different items, are inherently ordinal. They represent a rank order of performance but say nothing about the magnitude of difference between individuals. The subsequent transformation of these scores into norm-referenced scaled scores, using age-based percentiles, does not rectify this problem. It compounds it by embedding score meaning within a specific population context, thereby obliterating any hope of trait-invariant measurement.

There is no evidence that any subtest in the WAIS-4 battery conforms to the axioms of fundamental measurement. There is no demonstration of unidimensionality within subtests, let alone within broader indices. The various cognitive tasks, ranging from vocabulary definitions to digit span recall, matrix reasoning to symbol search, are not united by a single underlying latent construct. They engage diverse cognitive processes and are influenced by distinct neurological, cultural, and experiential factors. Aggregating these scores and treating them as if they reflect degrees of a unidimensional trait is entirely without psychometric foundation. More importantly, it prevents any transformation of scores into a scale where legitimate measurement claims, such as the possession of more or less of a cognitive attribute, could be made.

From the perspective of Rasch measurement theory, which is the only known framework capable of converting ordinal responses into ratio-level measurements of latent constructs, this entire enterprise collapses. Rasch provides a model in which person ability and item difficulty are placed on the same linear logit scale, grounded in probabilistic response theory. It is only through Rasch transformation that we can talk meaningfully about the amount of a trait an individual possesses, based on their pattern of responses. Rasch also requires that items function independently of the sample and that person measures are invariant across item sets. These properties are entirely absent in the construction and interpretation of WAIS-4 scores. No Rasch transformation is applied. No test of invariance is conducted. No model is estimated. Instead, scores are summed, normed, and narratively interpreted under the illusion that they represent quantities.

The profile-based approach, then, is not simply a soft form of interpretation; it is a form of clinical theatre. It allows psychologists to appear as though they are drawing scientifically grounded inferences from complex cognitive profiles when, in reality, they are applying pattern recognition to non-measurement artifacts. This is not to say that such interpretations have no descriptive or practical utility. In some contexts, profile differences may align with observable functional limitations. But their utility does not confer scientific legitimacy. Without meeting the axioms of measurement, such interpretations are not measures of trait possession. They are stories told about behavior, packaged as psychometrics.

The failure to recognize the fundamental limitations of WAIS-4 profile scores stems from a broader ignorance within the psychological community regarding what it means to measure something. This is not a mere technical oversight; it is a foundational error that renders claims about intelligence, memory, or cognitive speed epistemologically meaningless when derived from such instruments. No amount of clinical training, statistical software, or elaborate manuals can substitute for the absence of a valid measurement model. Unless psychologists are willing to engage with the demands of Rasch and build instruments that reflect the structure of latent trait possession, their claims remain trapped in a pre-scientific domain.

CASE STUDY 1: WAIS-4 DIGITAL SPAN

Consider the Digit Span subtest, a core element of the Working Memory Index in WAIS-4, as a concrete example. In Digit Span, the respondent is required to repeat sequences of numbers forwards, backwards, or in ascending order. Scores are derived by summing the number of correctly recalled sequences, producing a raw score that is then converted into a scaled score using

age-normed tables. This score contributes to the broader Working Memory Index and, ultimately, to the FSIQ.

First, the unidimensionality assumption is violated. Rasch measurement requires that all items in an instrument reflect a single latent trait. In Digit Span, however, the cognitive demands of the "forward," "backward," and "sequencing" tasks differ significantly. Forward digit span may rely primarily on short-term memory and attention, while backward span introduces executive functioning, and sequencing requires mental manipulation and ordering strategies. These are not manifestations of a single latent construct but different cognitive processes. Rasch diagnostics, such as principal component analysis of residuals, would likely show multidimensionality.

Second, the additivity of ordinal scores is assumed without justification. Summing the number of correctly recalled digit sequences produces an ordinal score, not a linear interval measure. These summed scores reflect a rank order of performance but do not support the inference of "how much more" one person possesses the latent trait than another. The step from correct/incorrect responses to summed raw scores to age-normed scaled scores masks the lack of measurement: the score transformations remain ordinal at best and fail the axioms of fundamental measurement.

Third, the WAIS-4 offers no recognition of the core principle of invariance that underpins Rasch measurement. In a Rasch framework, a person's ability estimate must be independent of the specific items administered, and item difficulty must remain constant regardless of the sample used. The WAIS-4 violates both conditions. Subtests such as Digit Span are normed within narrow age bands, but there is no evidence that item difficulties are stable across different populations or that person scores are invariant to item selection. The test's scoring depends entirely on where a respondent's raw score ranks within an age-specific reference group. This reliance on normative ranking, essentially a form of relative standing, is the opposite of Rasch's objective measurement model, which yields sample-independent, trait-level estimates based on an invariant metric. The WAIS, in effect, substitutes population-relative norms for genuine measurement, offering no assurance that scores reflect any stable underlying attribute across persons or items.

Fourth, no logit scale is constructed. Without placing persons and items on a common linear logit scale through joint maximum likelihood estimation or equivalent, the interpretation of "possession of working memory ability" becomes a narrative, not a measurement claim. Rasch transforms raw ordinal data into a meaningful quantitative metric that reflects the odds of a correct response, allowing for statements about trait possession. WAIS does nothing of the sort.

In sum, Digit Span, like the other WAIS subtests, does not support statements about the degree to which an individual possesses the latent trait of working memory capacity. It offers a norm-referenced score, not a scientifically valid measure. The failure to meet Rasch standards, unidimensionality, invariance, linearity, and the construction of a trait-centered logit scale, makes any claim about the extent of trait possession invalid. The subtest is thus diagnostic only in a loose, descriptive sense; it does not constitute measurement in the scientific sense demanded by Rasch theory.

CASE STUDY 2: WAIS-4 VOCABULARY SUBTEST

The vocabulary subtest of the WAIS-4, is often regarded as a core measure of verbal comprehension and crystallized intelligence. It is commonly assumed to be a proxy for verbal ability or accumulated knowledge. In the Vocabulary subtest, examinees are asked to define a series of increasingly difficult words. Each response is scored on a 0–2 scale: 0 for an incorrect or irrelevant response, 1 for a partially correct or vague definition, and 2 for a precise or textbook answer. These scores are summed and then transformed into a scaled score based on age-specific norms.

From a Rasch measurement perspective, this scoring approach is irredeemably flawed for several reasons. First, and most critically, the scoring scale is ordinal, not interval. The difference between a score of 0 and 1 is not equivalent in meaning or trait expression to the difference between a 1 and a 2. There is no demonstration that the underlying latent trait of "verbal comprehension" is manifested with equal intervals across the scoring categories. Rasch measurement requires ordered response thresholds that must be empirically validated; yet the WAIS Vocabulary subtest imposes these thresholds without justification. The assumption of equal distances between score levels is an illusion.

Second, the aggregation of item scores assumes additivity across non-equivalent items. In Rasch terms, each item must be calibrated for difficulty, and person measures should be located on the same logit scale as item difficulty. WAIS does not calibrate the Vocabulary items in this way. Instead, raw summed scores are assumed to reflect more or less of the trait. But this summation of item scores presumes that all items contribute equally to the latent variable, regardless of their differential difficulty or discrimination; an assumption which Rasch modeling explicitly tests and often rejects.

Third, no evidence of unidimensionality is provided. Vocabulary tasks may seem superficially unidimensional, but they actually engage a range of abilities: semantic knowledge, linguistic precision, socio-cultural exposure, even expressive fluency. Unless unidimensionality is tested, e.g., via Rasch residual principal components analysis or bifactor modeling, no claim can be made that the Vocabulary subtest is measuring a single latent construct. Without this, any resulting scale lacks interpretive clarity and violates the Rasch requirement for a single underlying dimension.

Fourth, the normative approach again fails the Rasch demand for invariance. In Rasch, person measures must not depend on the sample, and item difficulties must remain invariant across subgroups. WAIS violates both. Its norm-referenced scoring system yields different scaled scores for the same raw performance depending on age group norms. This means the same observed response pattern may imply different “trait levels” depending on which normative table is used—destroying any claim to sample-independent measurement.

Fifth, no transformation to a logit scale is attempted or reported. As a result, we have no meaningful, interval-level metric to represent possession of the latent verbal trait. Instead, we are left with raw sums and percentile ranks, neither of which supports mathematical operations nor empirical evaluation of change over time.

In conclusion, the WAIS Vocabulary subtest exemplifies the methodological failure of traditional IQ testing when viewed through the Rasch lens. Despite its apparent sophistication, the subtest cannot support any scientific claim regarding the possession of a latent verbal trait. Like the rest of the WAIS-4 battery, it confuses rank ordering with measurement and mistake norm-based scoring for trait quantification. It remains a subjective rubric-driven evaluation, not a tool for scientific inference.

IMPLICATIONS OF FAILED MEASUREMENT

The continued use of WAIS-4 profile scores, particularly in settings where decisions with significant consequences are made, clinical diagnosis, legal judgments, educational accommodations, represents a failure of both professional ethics and scientific rigor. It is the perpetuation of numerical storytelling in the guise of psychological expertise. Until profile-based interpretation is abandoned in favor of empirically grounded measurement, Australian psychology will continue to mistake score patterns for measurement, and inference for science.

This failure carries significant implications. In the absence of valid measurement, clinical judgments become vulnerable to misrepresentation, programs of support and intervention are misallocated, and legal determinations may be founded on psychometric illusions. What appears to be precise and objective is, under scrutiny, nothing more than a norm-referenced ordinal estimation based on flawed assumptions of scale and structure. The interpretation of subtest scatter, index discrepancies, or deviations from normative expectations lacks any epistemological legitimacy. These artifacts are not indices of latent trait possession; they are summaries of behavior coded into arbitrary score ranges and falsely elevated into psychological indicators.

In clinical practice, a diagnosis may depend on identifying whether an individual meets certain cognitive thresholds. These thresholds are defined numerically, typically by scaled scores or FSIQ without reference to the actual measurement structure of the trait under examination. For instance, an FSIQ below 70 may be used to identify intellectual disability, which in turn could determine eligibility for the NDIS or school-based special education services. Yet the number itself has no measurement status. It is a norm-based rank within an age band, not a trait-based measure. Rasch measurement makes this clear: unless the underlying construct is unidimensional, the response structure invariant, and the output placed on a logit scale, the resulting number cannot be interpreted as indicating "more" or "less" of a trait. Clinical judgment is then suspended upon nothing more than statistical folklore.

This becomes particularly critical in the context of program eligibility and exclusion. A misclassified IQ score can deny services to someone in need, or conversely, allocate scarce resources to someone for whom the classification is inapplicable. The magnitude of the error is obscured by the formal presentation of test scores. When a psychologist asserts that one child's verbal comprehension is "significantly below average" and another's processing speed is "within the normal range," the client and institutional stakeholders are led to believe that such categories are built on robust scientific foundations. But if those indices are constructed from ordinal scales without transformation to a measurement model, then they convey no valid difference in trait possession. No meaningful inferences can be made about developmental deficits, intervention

response, or expected academic functioning. The entire enterprise is built upon the manipulation of numbers that are not, in any fundamental sense, measures.

In legal contexts, the misuse of profile scores is even more consequential. An individual's cognitive profile may be introduced as evidence in court to determine criminal culpability, fitness to stand trial, or civil competence. Legal actors, judges, lawyers, jurors, rarely have the expertise to challenge the evidentiary basis of psychological assessments. They rely on the authority of the clinician and the appearance of numerical precision. But when the WAIS-4 is invoked to establish diminished responsibility or to support claims of cognitive impairment, the entire claim is anchored to a pseudo-measure. If the subtests used to infer impairment lack interval or ratio properties, then any resulting legal decision is contaminated by measurement failure. Courts may be persuaded by the appearance of expert analysis while never realizing that the conclusions are epistemically bankrupt.

Rasch measurement presents a profound challenge to this status quo. It enables the construction of instruments that do not merely rank responses but transform them into linear, invariant measures of latent trait possession. Rasch-conforming instruments demonstrate unidimensionality, sample and item invariance, and provide additive scales with interpretable logits. In the Rasch framework, we can make defensible statements about how much of a trait someone possesses and how that possession compares across persons and items. Such measurement makes it possible to challenge WAIS-4-based conclusions on methodological grounds. If the trait of interest, whether working memory, verbal comprehension, or processing speed, is not measured on an invariant linear scale, then claims derived from WAIS-4 data must be dismissed as descriptive fictions rather than scientific findings.

The availability of Rasch-based alternatives since the 1960s opens the door to forensic critique. A lawyer, for example, might ask whether a psychological opinion is based on a measure that satisfies the axioms of fundamental measurement. If not, then what basis exists for asserting that an individual's cognitive performance meets or fails to meet a legal threshold? Similarly, a government agency tasked with determining benefit eligibility based on IQ might be forced to reconsider its procedures. It cannot, in good faith, predicate access to life-changing programs on scales that lack any claim to ratio properties. As Rasch instruments proliferate, and as their methodological superiority becomes more widely understood, institutions will face increasing pressure to defend their reliance on antiquated ordinal tools.

The implications also extend to professional training and regulation. If Australian universities continue to teach psychometric interpretation without instruction in the axioms of measurement theory and the Rasch model, they are producing practitioners who unwittingly commit epistemic fraud. Regulators such as AHPRA, by maintaining silence on these matters, effectively sanction the continuation of scientifically invalid practices. The remedy is not incremental reform. It requires a paradigm shift in how psychologists are trained, certified, and held accountable for their use of instruments that purport to measure human traits.

In sum, the misuse of profile scores is not a harmless tradition. It has profound consequences for clinical judgment, program inclusion, and legal interpretation. It perpetuates an illusion of measurement where none exists and sustains a professional discourse that privileges appearance

over validity. Rasch measurement offers the only empirically and theoretically grounded alternative; an architecture for building instruments that can support claims about trait possession. Until Australian psychology embraces this alternative, the consequences of failed measurement will continue to ripple outward, distorting lives, misallocating resources, and undermining the very credibility of the discipline.

In sum, the misuse of profile scores is not a harmless tradition. It has profound consequences for clinical judgment, program inclusion, and legal interpretation. It perpetuates an illusion of measurement where none exists and sustains a professional discourse that privileges appearance over validity. Rasch measurement offers the only empirically and theoretically grounded alternative—an architecture for building instruments that can support claims about trait possession. Until Australian psychology embraces this alternative, the consequences of failed measurement will continue to ripple outward, distorting lives, misallocating resources, and undermining the very credibility of the discipline.

TOWARD A RASCH UNIVERSE: RECLAIMING MEASUREMENT IN CLINICAL PRACTICE

To move toward a Rasch universe, where psychological attributes are not merely observed but measured, a fundamental transformation must occur across the educational, institutional, and regulatory structures that currently underpin psychological practice in Australia. The existing framework, which uncritically accepts ordinal instruments and treats raw scores and normed percentiles as proxies for latent traits, must be replaced by a model grounded in the axioms of fundamental measurement. Rasch measurement is not a statistical refinement or psychometric embellishment; it is a necessary epistemic framework for transforming observations into measures that can support evaluable claims. The challenge now is for professional bodies, such as the APS, the AHPRA, and academic psychology departments, to assume the responsibility for this transition.

The first step must be educational reform. No psychologist should graduate or be licensed without a thorough understanding of what it means to measure a latent construct. At present, psychometric instruction is limited to legacy test theory, where the emphasis is on internal consistency, factor analysis, and normative interpretation. These approaches, which dominate the psychological testing canon, never confront the foundational question: is this a measure or merely a scale? The principles of Rasch measurement must be embedded in graduate education, not as a niche alternative, but as the standard by which all instruments are judged. Students must be taught that unless a latent construct is demonstrated to be unidimensional, its items to be invariant, and its scores to reside on a ratio scale, then no claim to trait possession can be justified. This reconceptualization of training will produce a new generation of psychologists equipped to design, critique, and defend instruments within a scientific framework.

Second, professional bodies such as APS must shift from passive endorsement of commercial instruments to active advocacy for scientifically valid measurement. APS cannot continue to certify or recommend tools like the WAIS-4 (or WAIS-5) without simultaneously addressing their epistemological failures. Instead, it must support the development and promotion of Rasch-conforming instruments and, crucially, provide clinicians with guidance on interpreting and using

these tools. Advocacy must extend to public and institutional stakeholders. Health services, courts, schools, and government agencies must be made aware of the limitations of existing instruments and the advantages of Rasch-based alternatives. APS must take the lead in re-educating practitioners and policymakers alike, issuing position statements that delineate the difference between ordinal scores and interval measures, and calling for the replacement of legacy tools with empirically grounded instruments.

Third, the clinical community must begin to construct a new profile instrument grounded in Rasch principles. This does not mean retrofitting the WAIS-4 (or the WAIS-5) to Rasch; it means abandoning it altogether. The Rasch model requires that the latent construct be defined clearly and that items be selected to represent increasing levels of trait difficulty along a single dimension. For clinical purposes, a new profile measure must be constructed for each trait of interest: verbal ability, working memory, cognitive speed, etc. Each profile would be generated not by aggregating unrelated subtests but by selecting items that conform to a Rasch hierarchy, empirically validated for difficulty and discrimination, and calibrated against a well-defined latent variable.

The construction of such instruments requires rigorous item development, trial administration across relevant populations, and iterative Rasch analysis. Items that fail to fit the model must be discarded. Instruments that show multidimensionality must be partitioned or restructured. The goal is not a test battery that mimics the appearance of WAIS-4 but one that builds linear, unidimensional, trait-specific profiles that support meaningful inference about a person's possession of the latent trait of interest. These measures must be anchored in invariant comparisons. A score must reflect the same level of ability across age, sex, and cultural background; or because a norm table says so, but because the measure itself is constructed to be invariant.

Once developed, these Rasch profile measures can be integrated into clinical decision-making. Instead of generating a FSIQ, the psychologist would present a Rasch logit-based trait profile, with each trait located on an interpretable continuum. Possession of a trait is not inferred from percentiles but directly estimated from the response pattern. Claims about ability are now statements about location on a linear scale. This enables both cross-sectional comparison and longitudinal tracking, as the same logit framework supports analysis over time or between individuals.

Importantly, Rasch profiles offer not only measurement integrity but also transparency. The method of estimation, the function of each item, and the assumptions of the model are all open to scrutiny. This makes them defensible in court, in education settings, and in healthcare. When challenged, psychologists can point not to norm tables or test manuals, but to the mathematical and empirical structure of the instrument itself. This reclaims psychology's authority not by institutional decree, but by adherence to the axioms of fundamental measurement.

Finally, AHPRA must revise its standards of professional conduct to reflect the imperative of scientific measurement. It is not ethically defensible to permit psychologists to base critical judgments on instruments that violate the axioms of measurement. If a diagnostic claim or eligibility determination is based on a non-measure, then that claim is professionally invalid. The regulatory framework must begin to demand evidence of measurement validity, including Rasch

conformity, for any instrument used in professional practice. This is not an abstract standard; it is a safeguard against error, injustice, and misallocated resources.

In summary, the transformation to a Rasch universe requires coordinated action across education, advocacy, instrument development, and regulatory oversight. It demands that psychologists abandon the comfortable fictions of ordinal psychometrics and instead commit to the rigors of real measurement. Only then can the profession restore its epistemic credibility and provide individuals with diagnoses and profiles that reflect not just observed behavior, but measured trait possession.

CONCLUSIONS

Australia must confront the reality that the WAIS-4 (and the presumptive WAIS-5), despite its institutional entrenchment, does not and cannot provide a valid measure of intelligence. Its outputs are ordinal, norm-referenced scores that masquerade as interval measures. These scores do not reflect possession of a latent trait but simply rank individuals relative to a population. In high-stakes settings, clinical, educational, vocational, and legal, this illusion of measurement leads to distorted decisions and unjustified outcomes. Profile-based interpretation, widely taught and practiced in Australia, merely compounds the problem by drawing inferences from score patterns that have no scientific foundation.

A shift toward Rasch-based measurement is imperative. Rasch provides the only epistemically valid method for constructing instruments that quantify latent traits, grounded in unidimensionality, invariance, and interval-level scaling. A new cognitive profile instrument designed specifically for Australian contexts must replace the WAIS-4. Such a measure would support meaningful, replicable claims about trait possession and restore integrity to psychological assessment. This is not a technical refinement but an ethical and scientific necessity. Continued use of WAIS-4 is an act of professional negligence. Australia must reject it and lead in the development of tools that align with the principles of real measurement.

ACKNOWLEDGMENT:

This paper benefited from the use of OpenAI's ChatGPT (GPT-4, 2024–2025), which served as a valuable tool in refining arguments, structuring content, and improving the clarity of presentation. While the author retains sole responsibility for all content, interpretations, and conclusions, the interactive support provided by this AI platform contributed meaningfully to the writing and editorial process

REFERENCES

[1] Wechsler D. Wechsler Adult Intelligence Scale – Fourth Edition (WAIS–IV). San Antonio, TX: Pearson Assessment, 2008

.

[2] Wechsler. D. WAIS–IV*:* Wechsler Adult Intelligence Scale – Fourth Edition (Australian Standard). Sydney: Pearson Clinical and Talent Assessment, 2008

[3] Roberts R., Goldacre. L. . Use of the Wechsler Adult Intelligence Scale (WAIS) in Australian clinical and forensic settings. *Australian Psychologist*, 2014; 49(5): 285–294

[4] Australian Psychological Society. Guidelines for the use of psychological tests. Melbourne: APS, 2017

[5] Bond T, Yan Z, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences (4th Ed). New York: Routledge, 2021

[6] Winter E, Tudel S. Wait, Where's the Flynn Effect on the WAIS-5. *J Intell*. 2024; 12(11): 118

[7] Wright B, Linacre J. Observations are always ordinal; measurement , however, must be interval. *Arch Phys Meas Rehab*. 1989; 70(12);857-860

[8] Rasch G. Probabilistic Models for some Intelligence and Attainment Tests (Expanded Ed). Chicago: University of Chicago Press, 1980 [originally published 1960]

[9] Krantz D, Luce R, Suppes P, Tversky A**.** *Foundations of Measurement*, Volumes I–III. New York: Academic Press, 1971 (Vol. I); 1989 (Vol. II); 1990 (Vol. III).

[10] Stevens S. On the theory of scales of measurement. *Science*. 1946;103:677-80

[11] Luce R, Tukey J. Simultaneous Measurement: A new type of fundamental measurement. *J Math Psychol*. 1964;1(1):1-27