AbilityScore: Development and Preliminary Psychometric Characterization of a Standardized Composite Index for Quantifying Child Developmental Readiness

Koti Reddy Saripalli[1,*], Raghupathi Ramtenki[2], Srikanth Yadav Boini[3]

[1]Bharath Healthcare Laboratories Private Limited, Hyderabad, Telangana, India

[2]Department of Pediatrics, TRR Institute of Medical Sciences, Patancheru, Hyderabad, Telangana, India

[3]Department of Neurology, Omega PRK Hospital, Chandanagar, Hyderabad, Telangana, India

[*]Corresponding author: kotii@kgvpl.com

March 2026

Technical Monograph v3.0

Abstract

Background: Since the late nineteenth century, developmental assessment has relied on a series of siloed instruments rather than a single developmental score. A child assessed with ADOS-2, Vineland and Bayley receives three separate reports on three incompatible scales, leaving families and clinicians without a unified picture of progress.

Objective: To describe the architecture, computation methodology, and preliminary psychometric evidence for AbilityScore, a standardized composite developmental index (0–1000) that quantifies observed functional developmental performance relative to age-normed expectations in multiple validated constructs.

Methods: AbilityScore synthesizes the clinical essence of 25 internationally validated assessments into a single metric, mapping 339 skills across 148 abilities through structured clinical observations. The scoring algorithm uses a Skill Index (60%) and Ability Index (40%) weighted composite with age-gender normalization. Seven domain-specific Readiness Indexes (Speech, Motor, Behaviour, Cognitive, School, Self-Sufficiency, Mainstream) provide multidimensional profiling. Preliminary psychometric evidence comes from operational data from approximately 18.6 million therapy sessions spanning more than 70 centers, 4 Indian states, 16 languages, and 28 diagnostic categories.

Results: Operational data show inter-rater reliability $r = 0.91$ ($p < 0.001$), intraclass correlation coefficient ICC = 0.88 (95% CI 0.84–0.91), mean absolute percentage error MAPE = 6.7%, and bias variance ratio ≤ 1.05. These real-world scoring consistency metrics from routine operations across more than 70 centers provide a robust operational foundation for the formal validation study now in planning, but do not substitute for controlled psychometric validation.

Conclusions: AbilityScore is a scientifically structured, WHO ICF-aligned composite index for developmental measurement. Formal concurrent validity testing against Vineland-3, CARS-2, and Bayley-4, along with test-retest reliability studies, are now in planning (BHCL-VAL-2026-001). This monograph documents the instrument in sufficient detail to support the upcoming validation studies, regulatory submissions, and clinical use.

Keywords: developmental assessment; composite index; child development; neurodevelopment; autism; WHO ICF; real-world evidence; psychometrics; AI-driven measurement

1 Introduction

1.1 The Problem: 160 Years Without a Unified Metric

For more than 160 years, developmental assessments have cataloged what children cannot do. The field relies on fragmented, single-domain tools: ADOS–2 for autism screening, Vineland for adaptive behavior, Bayley for infant development, CARS-2 for autism severity, Conners-3 for attention – each producing scores in different units, at different scales, administered by different professionals [Sparrow et al., 2016, Schopler et al., 2010, Bayley and Aylward, 2019].

In day-to-day practice at our centers, the same child may receive separate ADOS-2, Vineland, and Bayley reports that do not talk to each other, leaving parents unsure how to interpret progress. Therapists cannot track cross-domain gains. Governments cannot measure population-level readiness. Researchers cannot compare outcomes in all settings. This lack of a single developmental score makes it difficult to compare the results or design population-level programs [Lancet, 2017].

1.2 The Solution: AbilityScore

AbilityScore inverts the legacy model. Instead of cataloging deficits, it measures what children can do and identifies which abilities can be systematically grown. It does this by extracting the clinical essence of 25 established assessments, not replacing them, but unifying their information yield into a single, reproducible 0–1000 composite index.

Informally, we often explain AbilityScore to parents as playing a role similar to HbA1c in diabetes, a number that summarizes many signals, even though the underlying measurement is very different. HbA1c measures the biochemical ground truth; AbilityScore is a constructed composite index derived from multi-domain validated constructs. The comparison is to function and utility, not to the underlying measurement ontology.

1.3 What Makes AbilityScore Different

Every existing developmental assessment is a periodic snapshot – administered at intervals, scored in isolation, and filed. In our deployment, AbilityScore is not a stand-alone test. It is part of GPT-OS®, the operating system that runs our therapy workflow [Saripalli, 2025]. Table 1 summarizes the key differences.

2 Instrument Architecture

2.1 Three-Layer Taxonomy

AbilityScore operates through a three-layer hierarchical taxonomy:

Layer 1: Clinical Observations (more than 591). Each observation is a discrete, clinically defined behavior mapped to WHO ICF functional codes (b110–b760) and ICD-11 structural codes [WHO, 2001, 2019]. Each observation is scored within a defined range and classified into a three-zone system: Red (support needed), Yellow (developing), and green (age-appropriate).

Layer 2: Skills (339 unique across both age bands); Skills are measurable actions that express a developmental function. The taxonomy contains two models calibrated by age (Table 2).

Table 2: AbilityScore taxonomy structure by age band

Dimension	Toddler (0–4 yr)	Child (>4 yr)
Unique skills	129	252
Unique abilities	80	75
Source assessments	24 instruments	25 instruments
Developmental domains	9 primary	9 primary
Readiness Index weights	7 per skill	7 per skill
Safety flags	20 types	20 types

Taxonomy Versioning. The taxonomy has evolved across development cycles. Table 3 provides traceability.

Layer 3: Abilities (148 unique combined). Abilities describe broad functional domains. Each is scored 0–1000 using weighted skill averages, with weights derived from real-world therapy outcome data. The architecture mirrors the WHO ICF structure and the anatomical codes of ICD-11.

2.2 The 25 Source Assessments

AbilityScore extracts the clinical essence of 25 internationally validated assessment families: DP-4, ADOS-2/ADOS-T, ADI-R, ISAA, GARS-3, CARS-2, Conners-3, VABS-3, ABAS-3, PEDI, BASC-3, SSIS, Sensory Profile-2, SRS-2, PSI, FES, PLS-5, CELF-P2, NEPSY-II, CDI, M-FUN, Bayley-4, BOT-2, WPPSI-IV, and BRIEF-2. Each is mapped to ICF function domains and ICHI intervention codes [WHO, 2023].

2.3 Operational Dataset Provenance

The taxonomy weights and calibration parameters derive from a structured clinical dataset comprising approximately 495,000 children, more than 21 million therapy sessions, and approximately 2.2 billion data points collected across more than 70 centers in 4 Indian states (Telangana, Andhra Pradesh, Karnataka, Delhi), in 16 or more languages, spanning 28 ICD-10 diagnostic categories and an age range of 7 months to 16 years. The median follow-up is approximately 8 assessment cycles per child. Data completeness exceeds 92%. Quarterly bias audits maintain variance ratio ≤ 1.05. The infrastructure is certified according to ISO/IEC 27001: 2022.

3 Computation Methodology

3.1 Scoring Formula

The composite AbilityScore is calculated using a dual-index weighted formula with age-gender normalization [Saripalli, 2025]:

$$\text{AbilityScore} = [\text{SI} \times 0.6 + \text{AI} \times 0.4] \times \text{AGF} \quad (1)$$

where SI is the Skill Index (0–100), AI is the Ability Index (0–100), and AGF is the Age-Gender Adjustment Factor.

Skill Index (60% weight): Aggregate weighted individual skill metrics. Each of the 339 skills receives an impact factor weight derived from multivariate gradient-boosted regression, with therapist-rated functional improvement delta as the dependent variable. Feature importance scores determine weights.

Ability Index (40% weight): Groups skills into developmental domains and computes a hierarchical domain-weighted composite using the same regression framework.

Age-Gender Adjustment Factor: Divides actual performance by expected performance for the child's demographic cohort:

Table 1: Comparison of AbilityScore with conventional developmental assessments

Dimension	Conventional Assessments	AbilityScore
Function	Periodic snapshot	Living measurement loop
Frequency	Annual or ad hoc	Monthly (or custom interval)
Administered by	Professionals in clinics	Parents, therapists, caregivers (AI-supported)
Output	Single-domain score	Universal 0–1000 + 7 Readiness Indexes
Cross-domain integration	None (siloed tools)	Unified across all domains
Intervention linkage	None	Direct: score drives therapy planning
Feedback loop	None	Human-in-the-Loop rating (1–9 stars daily)
Population comparability	Impossible across tools	Standardized metric (cross-population validation pending)
Scale	Dozens to hundreds	Approximately 495,000 children, 21M sessions

Table 3: AbilityScore taxonomy version history

Version	Skills	Abilities	Observations	Context
v1.0 (Monograph)	339	148	>591	Combined taxonomy (canonical)
v0.9 (Deployment)	344	79	529	Operational subset
v0.8 (Patent)	333	74	508	IP reference

$$\text{AGF} = \frac{\text{Observed Performance}}{\text{Expected Performance}_{(\text{age, gender})}} \quad (2)$$

3.1.1 Dataset Partition for Weight Derivation

To avoid optimistic bias, the operational data set is partitioned: Training set (60%) for weight derivation; Validation set (20%) for cross-validation; Holdout set (20%) for independent evaluation, never used in model building. Partitioning is stratified by diagnosis, age band, sex, and center.

3.1.2 Age-Gender Normalization

The reference population derives from composite normative data embedded in the 25 source assessments (typically n = 1,000–4,000 per instrument). Normalization uses a Z-score with min-max scaling. The calibration constants (k_i, p_i) are locked under the ISO 13485 change control.

3.2 Zone Classification

Zone boundaries were established through outcome-based clustering analysis of therapy response trajectories in the operational data set. Specifically, clustering of k-means in baseline-to-outcome trajectories identified natural inflection points in the probability of clinically significant improvement, which were then validated by multidisciplinary clinical panels in more than 70 centers. The 9-grade system ranges from Critical Red (0–110) to Elite Green (881–1000), with each zone associated with specific clinical meaning and expected improvement timelines.

3.3 Seven Readiness Indexes

Each skill carries seven Readiness Index weights, enabling domain-specific composites: Speech ($\text{RI}_{\text{Speech}}$), Motor ($\text{RI}_{\text{Motor}}$), Behaviour ($\text{RI}_{\text{Behaviour}}$), Cognitive ($\text{RI}_{\text{Cognitive}}$), School ($\text{RI}_{\text{School}}$), Self-Sufficiency ($\text{RI}_{\text{SelfSuff}}$), and Mainstream ($\text{RI}_{\text{Mainstream}}$). Each is scored 0–100.

4 The GPT-OS Lifecycle

AbilityScore operates within GPT-OS® as a six-step closed loop:

1. **MEASURE:** Initial 90-minute structured assessment that produces AbilityScore, 7 Readiness Indexes, and a Red/Yellow/Green zone map.
2. **INSTANTIATE:** Personal Development Kernel (PDK) created, the child's living digital twin (400+ data points).
3. **ACT:** TherapeuticAI generates a personalized Everyday Therapy Program (7–9 techniques per domain, delivered 4 days ahead). In a parent-assisted, therapist-supervised model, families implement 3 techniques daily.
4. **FEEDBACK:** Human-in-the-Loop Fusion Module—parents rate each technique 1–9 stars. PDK recalibrates daily.
5. **REFRESH:** Monthly reassessment from previous baseline. Delta computation tracks improvement trajectory.
6. **SCALE:** Same system operating across more than 70 centers (approximately 1,600 therapists, more than 12,000 children daily) and extending into homes.

5 Operational Performance Metrics

The following data come from routine clinical operations in more than 70 Pinnacle Blooms Network centers – not from a controlled psychometric validation study. We present them as evidence of the consistency of the real-world scoring, not as a substitute for formal validation (Table 4).

5.1 Content Validity

The validity of the content comes from the synthesis of 25 validated assessment instruments and 16 years of iterative clinical refinement. Taxonomy mapping and clinical construct harmonization were developed through multidisciplinary expert consensus involving developmental pediatricians (including RR, who reviewed the structure of the developmental domain and the calibration of the age-band with clinical practice in a tertiary pediatric teaching hospital), pediatric neurologists (including SYB, who reviewed the mapping of the neurological construct, the architecture of the safety flag and the exclusion of confounding neurological conditions), speech and language therapists, occupational therapists, behavioral analysts, and special educators in more than 70 centers of the Pinnacle Blooms Network.

5.2 Construct Validity

Supported by zone-severity correlation, Readiness Index-domain mapping, sensitivity to change (+150.8 points mean gain over 6 months, 83.33% Red Zone migration), and WHO ICF alignment.

5.2.1 Definition of Functional Improvement

Functional improvement is operationalized as a multidimensional construct comprising three independent criteria: (i) a positive therapist-rated functional gain on a structured domain-specific rating scale completed independently of AbilityScore; (ii) achievement of individualized therapy goal completion targets (percentage of SMART goals met per assessment cycle); and (iii) a zone transition or within-zone score increase exceeding the preliminary MCID threshold of 25 points. This multi-criteria definition prevents circular validation where the score defines improvement and improvement validates the score.

5.2.2 Reliability Does Not Imply Validity

The high inter-rater reliability ($r = 0.91$) tells us that the instrument produces consistent scores between raters and centers. But consistency is not accuracy. A well-calibrated thermometer that always reads 2 degrees too high is reliable but not valid. We need formal concurrent validity tests –comparing AbilityScore with Vineland-3, CARS-2, and Bayley-4 in the same children – before we can claim validity. That study is now in planning as BHCL-VAL-2026-001.

5.2.3 AI Role in Scoring

AbilityScore computation is *deterministic* once the weights are fixed. Given the same set of skill observations and the same age-gender cohort, the same score is always produced, regardless of when or where the computation runs. AI (specifically gradient-boosted regression) was used only during the weight derivation phase in the training set. At runtime, scoring applies pre-computed, version-locked weights under ISO 13485 change control, ensuring longitudinal comparability across assessment cycles. In short: the measurement layer is a fixed-weight composite model; the recommendation layer (TherapeuticAI) is the adaptive part.

5.2.4 Normative Data

Age-gender normalization is derived indirectly from the standardization samples published with the 25 source instruments. We have not yet collected our own normative cohort of typically developing children assessed directly with AbilityScore. A dedicated normative study (planned $n = 1{,}000+$, including children that are typically developing and neurodevelopmentally diverse in multiple Indian states) is in preparation.

5.3 Known Gaps

We are candid about what remains unproven. Concurrent validity, test-retest reliability under controlled conditions, and a formal MCID have not been published. A dedicated validation study (BHCL-VAL-2026-001, $n = 300$, cross-sectional design of the same-child, same-session with blinded comparator raters) is now in active planning. Until those results are available, AbilityScore should be regarded as a well-constructed instrument with strong operational evidence, not yet a formally validated one.

6 Standards Alignment

AbilityScore aligns with: WHO ICD-11 + ICF + ICHI [WHO, 2001, 2019, 2023]; WHO AI-for-Health Guidance [WHO, 2021]; UNICEF AI-for children [UNICEF, 2021]; ISO 13485:2016; ISO/IEC 27001:2022; ISO/IEC 23894:2023; IEEE P7003; HL7 FHIR R5; SNOMED CT; LOINC.

7 Intellectual Property

Indian Patent Application No. 202541049714 (Filed 23 May 2025): system + method + CRM claims. 13+ WIPO PCT applications across 150+ countries. CDSCO SaMD classification pending (Ref: MD/classification/2025/150).

Table 4: Preliminary operational psychometric metrics

Property	Result	Benchmark	Interpretation
Inter-rater reliability (r)	0.91 ($p < 0.001$)	≥ 0.85	Excellent
ICC	0.88 (95% CI 0.84–0.91)	≥ 0.75	Good–Excellent
MAPE	6.7%	$\leq 10\%$	Target met
Bias Variance Ratio	≤ 1.05	≤ 1.10	No significant bias
Latency	212 ms/child	≤ 300 ms	Real-time confirmed

8 Regulatory Classification

AbilityScore is a non-medical, non-diagnostic developmental measurement system. The regulatory architecture maintains hard separation: AbilityScore (non-diagnostic measurement), TherapeuticAI (clinical decision support), GPT-OS (supervised intervention platform).

9 Limitations

1. **Concurrent validity:** We have not yet published formal correlation coefficients against established instruments. The validation study (BHCL-VAL-2026-001) is designed to close this gap.

2. **Population scope:** All data come from Indian populations in 4 states and 16 languages. In particular, we have no data yet from high-income settings or from health systems with very different service delivery models, so we do not assume that the current parameters will generalize without recalibration.

3. **MCID:** Not formally established. Our preliminary estimate of 25–30 points per month comes from clinical observation, not from a controlled anchor-based study.

4. **Test-retest:** We have not conducted a formal test-retest study under controlled conditions (short interval, no intervening therapy) separately from our operational reliability data.

5. **Normative cohort:** Our age-gender norms are borrowed from the source instruments' standardization samples, not collected directly with AbilityScore. A dedicated normative study (n = 1,000+) is planned.

Ethical Considerations

This monograph describes instrument architecture and reports aggregate operational metrics derived from de-identified clinical data. Individual participant data are not presented. The operational data set is maintained under ISO/IEC 27001:2022 certified infrastructure in accordance with the India Digital Personal Data Protection Act 2023 and the EU General Data Protection Regulation. Formal validation studies referencing this monograph (BHCL-VAL-2026-001) will obtain the approval of the independent Ethics Committee prior to enrollment of participants.

Funding

This work was funded by Bharath Healthcare Laboratories Private Limited (BHCL). External funding was not received.

Conflicts of Interest

KRS is the Founder and Chairman of Bharath Healthcare Laboratories Private Limited, which developed and operates AbilityScore as part of the GPT-OS platform. This conflict is fully disclosed. RR contributed to the validation of the clinical construct and the review of the development domain as an independent academic pediatrician; SYB contributed to neurological construct mapping and safety flag architecture review as an independent consultant neurologist. Neither RR nor SYB holds equity, employment, or financial interest in BHCL. Formal validation studies employ an independent external Principal Investigator with sole authority over study conduct and publication decisions. The sponsor cannot veto, suppress, or delay publication of unfavorable findings from validation studies.

Data Availability

The aggregate operational metrics are reported in this monograph. Individual-level data are not publicly available due to privacy regulations. Requests for de-identified research datasets should be directed to the corresponding author and are subject to Ethics Committee approval and data use agreements.

AI Assistance Disclosure

I used AI-assisted tools to draft and edit the manuscript text. All scientific content, data, instrument design, clinical decisions, and final approval are my responsibility. The AI tools did not have access to the individual participant data.

References

WHO. International Classification of Functioning, Disability and Health (ICF). Geneva: World Health Organization; 2001 (updated 2017).

WHO. International Classification of Diseases, 11th Revision (ICD-11). Geneva: World Health Organization; 2019.

WHO. International Classification of Health Interventions (ICHI). Geneva: World Health Organization; 2023.

WHO. Ethics and Governance of Artificial Intelligence for Health. Geneva: World Health Organization; 2021.

UNICEF. Policy Guidance on AI for Children. New York: UNICEF; 2021.

Sparrow SS, Cicchetti DV, Saulnier CA. Vineland Adaptive Behavior Scales, Third Edition (VABS-3). San Antonio, TX: Pearson; 2016.

Schopler E, Van Bourgondien ME, Wellman GJ, Love SR. Childhood Autism Rating Scale, Second Edition (CARS-2). Los Angeles: Western Psychological Services; 2010.

Bayley N, Aylward GP. Bayley Scales of Infant and Toddler Development, Fourth Edition. San Antonio, TX: Pearson; 2019.

Lancet. Early childhood development coming of age: science through the life course. *Lancet*. 2017;389(10064):77–90.

Saripalli KR. A System for Assessing Developmental Readiness of a Child and Method Thereof. Indian Patent Application No. 202541049714. Filed 23 May 2025.

Saripalli KR. A Concurrency-Enabled, Accelerator-Based System and Method for Real-Time Autism Ability and Self-Sufficiency Scoring. Indian Patent Application. Filed December 2024.

Saripalli KR et al. A Physiological Computation Framework for Quantifying and Visualising Child Development: The GPT-OS Diagnosis Engine. Technical Report, Bharath Healthcare Laboratories; 2025.

Raina SK et al. Prevalence of autism spectrum disorder among rural, urban, and tribal children. *J Neurosci Rural Pract*. 2017;8(3):368–374.

Juneja M et al. Prevalence of autism spectrum disorders in school-aged children in India. *Indian Pediatr*. 2020;57(10):907–910.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310.

Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163.

ICMR. National Ethical Guidelines for Biomedical and Health Research Involving Human Participants. New Delhi: ICMR; 2017.