Concurrent Validity, Test-Retest Reliability, and Preliminary Minimal Clinically Important Difference of PinnacleAI AbilityScore: A Cross-Sectional Psychometric Validation Study Protocol

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

[†]Co-Investigator (Sponsor). Corresponding author: kotii@kgvpl.com

Clinical Research Consultancy: Manentia Research Pvt Ltd

March 2026

Study Protocol v2.1 FINAL | SPIRIT 2013 Compliant

Study ID: BHCL-VAL-2026-001 | Short Title: PINNACLE-VAL

Abstract

Background: AbilityScore is a standardized composite developmental index (0–1000) that quantifies child developmental performance relative to age-normed expectations. Operational data from more than 18.6 million therapy sessions show inter-rater reliability $r = 0.91$ and ICC $= 0.88$. However, formal concurrent validity against established reference instruments, controlled test-retest reliability, and minimal clinically important difference (MCID) have not been published.

Objective: To determine the concurrent validity of AbilityScore against Vineland-3, CARS-2, Bayley-4, and ABAS-3; to assess the reliability of the 7-day test-retest; and to estimate a preliminary MCID.

Methods: Cross-sectional multi-center psychometric validation study with a nested prospective test-retest substudy. Component A: $n = 300$ children aged 7 months to 7 years assessed with AbilityScore and comparator instruments in the same session by blinded raters in 6 Pinnacle Blooms Network centers. Component B: $n = 60$ subset, reassessed at 7 days. Primary analysis: Pearson correlation (Bonferroni-adjusted $\alpha = 0.017$). Pre-registered SAP on OSF.

Discussion: This study addresses the single most important evidence gap for AbilityScore. It will determine whether the instrument meets the accepted thresholds for concurrent validity ($r \geq 0.50$) and reliability (ICC ≥ 0.80), allowing the progression from operational evidence to formal psychometric validation.

Trial registration: CTRI (to be registered prior to first enrollment). OSF pre-registration of SAP (to be completed prior to data collection).

Keywords: psychometric validation; concurrent validity; test-retest reliability; MCID; developmental assessment; AbilityScore; SPIRIT 2013

1 Background and Rationale

AbilityScore is a composite developmental index that we built to address a practical problem: at our centers, the same child receives ADOS-2, Vineland and Bayley reports that do not talk to each other, and families are left unsure how to interpret progress. AbilityScore unifies the clinical essence of 25 established assessments into a single 0–1000 metric with seven domain-specific Readiness Indexes. The instrument architecture, the computation methodology, and the operational evidence are described in the Technical Monograph (Saripalli, 2026).

Operational data from approximately 18.6 million therapy sessions demonstrate inter-rater reliability $r = 0.91$ ($p < 0.001$) and ICC $= 0.88$ (95% CI 0.84–0.91). These real-world scoring consistency metrics from routine clinical operations in more than 70 centers provide a robust operational foundation for formal validation, but do not substitute for a controlled psychometric validation study.

Three critical properties remain unestablished through peer-reviewed evidence:

1. **Concurrent validity:** No formal correlation coefficients have been published between AbilityScore and established reference instruments administered to the same children in the same session.
2. **Reliability of the test and the retest:** No formal assessment of the stability of the score has been performed over a short period without intervening therapy modification.
3. **Minimal clinically important difference (MCID):** The smallest score change that families and therapists consider meaningful has not been formally determined.

This study addresses all three gaps. We are candid: until these results are available, AbilityScore should be regarded as a well-constructed instrument with strong operational evidence, not yet a formally validated one.

2 Objectives

2.1 Primary Objective

Determine the concurrent validity of AbilityScore by computing correlation coefficients between AbilityScore (composite and 7 Readiness Indexes) and established reference instruments (Vineland-3, CARS-2, Bayley-4, ABAS-3) when administered to the same children in the same assessment session.

2.2 Co-Primary Objective

Determine the reliability of AbilityScore test and retest over a 7-day interval (±2 days) without intervention therapy modification, assessed by ICC (mixed two-way, absolute agreement).

2.3 Secondary Objectives

1. To estimate a preliminary MCID using both anchor-based (therapist-rated Global Impression of Change) and distribution-based (0.5 SD, 1 SEM) methods.

2. To evaluate classification concordance of AbilityScore zones (Red/Yellow/Green) against severity categories from comparator instruments (exploratory ROC-based analysis—not diagnostic accuracy, as AbilityScore is explicitly non-diagnostic).

3. Assess agreement between AbilityScore Readyiness Indexes and corresponding comparator domain scores using Bland-Altman analysis.

2.4 Exploratory Objectives

1. Examine whether the AbilityScore baseline is associated with the duration of therapy at 6 months (pilot analysis generating hypotheses only).

2. To compare psychometric properties across diagnostic subgroups (ASD, ADHD, GDD, Speech-Language Disorder) – descriptive subgroup analysis.

2.5 Publication Hierarchy

This study supports two manuscripts. The primary manuscript will report concurrent validity and test-retest reliability (objectives 2.1 and 2.2). The secondary manuscript will report preliminary MCID, classification concordance, and subgroup analyzes (objectives 2.3 and 2.4). The primary manuscript is the priority delivery.

3 Study Design

Cross-sectional multi-center psychometric validation study with a nested prospective test-retest substudy (Table 1).

All 300 children complete Component A. A random subset of 60 returns 7 days later without any modification to their therapy programme.

3.1 Comparator Instruments

We selected comparator instruments that the global developmental assessment community treats as established references – not because they are perfect, but because they are the language that academia speaks (Table 2).

The coverage of the comparator is stronger in the Toddler band (three instruments covering adaptive, autism-specific, and developmental domains) than in the Child band (three instruments covering adaptive and autism-specific but lacking a direct broad cognitive comparator). We considered adding WPPSI-IV for the Child band but excluded it: children with neurodevelopmental conditions aged 4–7 show markedly elevated assessment fatigue with IQ-type testing, and we prioritized ecological validity and child tolerance over maximal domain coverage.

Total assessment burden: Toddler band approximately 150–250 minutes; Child band approximately 135–180 minutes. Sessions are split into two consecutive days with a mandatory 15-minute rest period between instruments. If a child shows sustained distress lasting more than 5 minutes or refuses to engage, the session is paused and resumed the following day.

4 Study Population

4.1 Inclusion Criteria

1. Age 7 months to 7 years of age at enrollment.

2. Documented neurodevelopmental condition (ASD, ADHD, GDD, Speech-Language Disorder, DCD, or Intellectual Disability).

3. Currently enrolled in a participating Pinnacle Blooms Network center.

4. Written informed consent from the parent or legal guardian.

For Component B (test-retest) and MCID estimation only, an additional criterion applies: at least one previous AbilityScore assessment in record, to enable the therapist Global Impression of Change rating. This requirement does *not* apply to Component A, where first-time AbilityScore participants are eligible to reduce familiarity bias.

4.2 Exclusion Criteria

1. Acute medical illness that could affect performance.

Table 1: Study design overview

Component	Design	Sample	Instruments	Primary Analysis
A: Concurrent validation	Cross-sectional; same child, same session	$n = 300$	AbilityScore + Vineland-3 + CARS-2 + Bayley-4 (toddler) / ABAS-3 (child)	Pearson/Spearman correlation, Bland-Altman, ROC
B: Test-retest	Prospective; 7-day interval	$n = 60$ (subset of A)	AbilityScore administered twice	ICC, Bland-Altman, SEM

Table 2: Comparator instruments by age band

Instrument	Domain	Rationale	Age band	Time
Vineland-3	Adaptive behavior	Most widely used adaptive measure; maps to Self-Sufficiency and Mainstream indexes	All (7mo–7yr)	20–60 min
CARS-2	Autism severity	Established severity classification; maps to Behaviour and Cognitive indexes	All (7mo–7yr)	5–10 min
Bayley-4	Cognitive, language, motor	Established infant/toddler assessment; domain scores comparable to composite	Toddler only (≤4yr)	30–90 min
ABAS-3	Adaptive skills	Broad functional comparator for child band where Bayley-4 is not applicable	Child only (>4–7yr)	15–20 min

2. Severe uncorrected sensory impairment that prevents valid comparator administration.

3. Parent unable to participate in the Vineland-3 caregiver interview.

4. Currently in another interventional research study.

5. The family has withdrawn their consent to the use of clinical data in research.

4.3 Sample Size

For concurrent validity: $n = 300$ provides 95% power to detect $r = 0.30$ at Bonferroni-adjusted $\alpha = 0.017$ and 99% power for $r = 0.50$. For test-retest: $n = 60$ provides 80% power to detect ICC = 0.80 with 95% CI width ≈ 0.15.

4.4 Sampling Strategy

Primary stratification (enforced): age band (Toddler/Child, 150 each) and diagnosis (ASD/ADHD/GDD/SLD, 75 each, minimum 50). Secondary (targeted): severity zone. Descriptive: sex, city tier. A CONSORT-style flow diagram will document participants screened, consented, completed, withdrawn (with reasons), and replaced.

5 Study Sites

Six Pinnacle Blooms Network centers selected for geographic and city-tier diversity: Gachibowli and Kukatpally (Hyderabad, Tier 1), Visakhapatnam and Vijayawada (Andhra Pradesh, Tier 2), Indira Nagar (Bengaluru, Tier 1), and Karimnagar (Telangana, Tier 2). Each site enrolls approximately 50 children.

All six sites are operated by sponsors. We acknowledge this openly: external site validation at non-Pinnacle centers is a necessary future step that this protocol does not address.

6 Procedures

6.1 Consent

Written informed consent from a parent or legal guardian in their preferred language (Telugu, Hindi, English, or Kannada). Unlike the main cohort study (BHCL-OBS-2026-001), which analyzes de-identified historical records, this study involves active participant contact and additional assessments, so individual informed consent is required. No waiver is sought.

6.2 Component A: Concurrent Validation

Within 48 hours of enrollment:

Session 1 (Day 1): AbilityScore administered by the child's assigned therapist (approximately 90 minutes). Immediately afterward, a *separate* clinician (blind to AbilityScore results) administers CARS-2 (approximately 10 minutes).

Session 2 (Day 1 or Day 2): A trained assessor (blinded to AbilityScore) administers Vineland-3 (caregiver interview, 20–60 minutes) and Bayley-4 (toddlers only, 30–90 minutes) or ABAS-3 (children >4 years, 15–20 minutes).

6.3 Component B: Test-Retest

Sixty children return 7 ± 2 days after Component A for a second AbilityScore evaluation only. During the interval: no changes in intensity, modality, frequency, or therapy goals. The actual therapy sessions attended during the interval are documented. The same therapist administers the retest where possible; different-therapist scenarios are documented and analyzed separately.

6.4 Preliminary MCID Estimation

In component A, the child's primary therapist completes a Global Impression of Change (GIC) rating on a 7-point Likert scale compared to the child's previous AbilityScore. This is a preliminary MCID estimate from a cross-sectional anchor, not a definitive MCID, which requires prospective longitudinal data in a separate study.

6.5 Blinding

All comparator raters are blinded to AbilityScore results. The AbilityScore therapist is blinded to comparator scores. Sealed envelopes until data lock.

6.6 Clinical Oversight Investigators

Two clinical oversight investigators provide independent medical governance for the study. They are not psychometric raters and do not administer AbilityScore or comparator instruments. Their defined responsibilities are as follows:

Dr. Raghupathi Ramtenki, MD Pediatrics (Associate Professor, Department of Pediatrics, TRR Institute of Medical Sciences, Hyderabad): (a) reviews and confirms the source diagnostic documentation for eligibility adjudication in ambiguous or complex cases; (b) provides clinical interpretation of developmental comparator findings (Vineland-3, Bayley-4, ABAS-3) when results diverge from AbilityScore zone classification; (c) advises on age-appropriate assessment modifications for children with comorbid medical conditions.

Dr. Srikanth Yadav Boini, MBBS MD (Gen Med) DM (Neurology) (Consultant Neurologist, Omega PRK Hospital, Chandanagar, Hyderabad; APMC Reg No. APMC/FMR/75069): (a) reviews eligibility for children with known or suspected neurological comorbidities (epilepsy, cerebral palsy, acquired brain injury) to confirm that confounding neurological instability does not compromise assessment validity; (b) provides neurological differential interpretation for complex cases where developmental regression or atypical progression is observed during the study; (c) supports protocol governance for safety-relevant neurological findings identified incidentally during comparator assessment.

7 Outcomes

Primary: Pearson r between AbilityScore composite and (a) VABS-3 Adaptive Composite, (b) CARS-2 Total, (c) Bayley-4 Composite (toddlers) or ABAS-3 GAC (children).

Co-primary: ICC between AbilityScore T1 and T2 at 7 days.

Secondary: Preliminary MCID (anchor + distribution); classification concordance (ROC-AUC of zones vs. comparator severity); Readiness Index agreement (Bland-Altman).

Exploratory: Predictive validity pilot; subgroup psychometrics.

Pre-specified directional hypotheses are shown in Table 3.

Strong correlation is evidence of convergent validity—it does not establish interchangeability between AbilityScore and the comparator instruments.

8 Statistical Analysis

Full details are in the pre-registered SAP (OSF). Key elements:

Populations: Full Analysis Set (completed AbilityScore + $\geq$1 comparator), Per-Protocol Set (no deviations), Test-Retest Set ($n = 60$), MCID Set (prior AbilityScore + GIC available).

Primary analysis: Pearson r with 95% CI (Fisher z). Bonferroni correction for 3 primary comparisons ($\alpha = 0.017$). Spearman ρ if normality violated.

Test-retest: ICC(3,1) with 95% CI. SEM = SD $\times$ $\sqrt{1 - \text{ICC}}$. Bland-Altman mean bias and limits of agreement.

MCID: Anchor-based (mean delta in GIC = 5 group), distribution-based (0.5 SD, 1 SEM). Triangulated as median of three estimates. Labeled "preliminary" throughout.

Classification concordance: ROC-AUC of zone boundaries vs. comparator severity categories. Exploratory only—not diagnostic accuracy.

Multiple comparisons: Bonferroni for primary family. Benjamini-Hochberg FDR for secondary. No adjustment for exploratory.

Missing data: Complete-case pairwise. No imputation. Withdrawn participants replaced to maintain $n = 300$.

Software: R 4.3+ (psych, pROC, BlandAltmanLeh, ggplot2, boot). Scripts on OSF.

9 Ethics and Governance

This study involves active participant contact and administration of assessments beyond routine care—specifically, comparator instruments (Vineland-3, CARS-2, Bayley-4, ABAS-3) administered for research validation. Individual informed consent is required.

Risk: Minimal. No experimental intervention. Only additional burden is comparator assessments.

Table 3: Pre-specified directional hypotheses and minimum effect thresholds

Comparison	Direction	Threshold	Rationale
AbilityScore vs VABS-3 Adaptive Composite	Positive	$r \geq 0.50$	Functional constructs overlap
AbilityScore vs CARS-2 Total	Negative	$r \leq -0.40$	Higher ability = lower severity
AbilityScore (Toddler) vs Bayley-4 Composite	Positive	$r \geq 0.45$	Cognitive-motor overlap
AbilityScore (Child) vs ABAS-3 GAC	Positive	$r \geq 0.50$	Adaptive function overlap

Fatigue mitigated by 2-day split and mandatory rest periods.

Conflict of interest: I (KRS) am the Founder and Chairman of the sponsor organization and Chief Architect of the platform that includes AbilityScore. Dr. Ramtenki (RR) and Dr. Boini (SYB) are independent clinical consultants with no equity, employment, or financial interest in BHCL. Five independence controls are in place: (1) external academic PI with no financial relationship with BHCL; (2) independent biostatistician executing the pre-registered SAP; (3) blinded comparator raters; (4) pre-registered SAP on OSF; (5) the sponsor cannot veto, suppress, or delay publication of unfavorable findings.

Data protection: Pseudonymized (UUIDv4), AES-256 encrypted, ISO/IEC 27001:2022 certified infrastructure. DPDP Act 2023 and GDPR compliant.

10 Timeline

1. **Months 1–2:** EC approval, CTRI registration, assessor training.
2. **Months 2–4:** Enrollment and Component A assessments ($n = 300$).
3. **Months 3–5:** Component B test-retest ($n = 60$).
4. **Month 5:** Data lock (SHA-256 hash).
5. **Months 5–6:** an independent biostatist executes SAP.
6. **Months 6–7:** Manuscript preparation and submission.

Total: 7 months from EC approval to manuscript submission.

Target journals: Lancet Digital Health (1st choice), JAMA Pediatrics (2nd), Journal of Child Psychology and Psychiatry (3rd).

11 Discussion

This protocol describes what we believe is the most important study in AbilityScore's evidence trajectory. If the results meet pre-specified thresholds – concurrent validity $r \geq 0.50$ against VABS-3 and test-retest ICC ≥ 0.80 – AbilityScore moves from a well-constructed instrument with strong operational evidence to a formally validated developmental metric. If results are weaker than expected, they are still valuable: they tell us exactly where the instrument needs refinement, and the pre-registered design ensures the findings are publishable regardless of direction.

We designed this study to be difficult to attack. Blinded raters prevent expectation bias. The hold-out data set prevents circular validation. The pre-registered SAP prevents post-hoc manipulation. The sponsor-cannot-veto clause prevents suppression. And the honest acknowledgment that all sites are sponsor-operated prevents the appearance of hidden conflicts.

What this study cannot do is establish international generalizability. All data comes from Indian populations in 4 states and 16 languages. Cross-cultural validation is the logical next step, planned for WHO/UNICEF pilot programs in 2026–2027.

Ethical Approval

To be obtained from a NAITIK-registered Ethics Committee prior to enrollment. CTRI registration before the first participant.

Funding

Funded by Bharath Healthcare Laboratories Private Limited. No external funding.

Conflicts of Interest

KRS is the Founder and Chairman of BHCL, which developed AbilityScore. RR and SYB are independent clinical consultants with no equity or financial interest in BHCL. Formal independence controls are described in Section 9. The sponsor cannot veto the publication of unfavorable findings.

Data Availability

De-identified study data will be made available upon reasonable request after manuscript publication, subject to Ethics Committee approval and data use agreements.

AI Assistance Disclosure

I used AI-assisted tools to draft and edit the protocol text. All scientific content, study design, statistical methods, and final approval are my responsibility. The AI tools did not have access to participant-level data.

References

Saripalli KR, Ramtenki R, Boini SYB. AbilityScore: Development and Preliminary Psychometric Characterization of a Standardized Composite Index for Quantifying Child Developmental Readiness. Technical Monograph v3.0. Bharath Healthcare Laboratories; March 2026.

Sparrow SS, Cicchetti DV, Saulnier CA. Vineland Adaptive Behavior Scales, Third Edition. Pearson; 2016.

Schopler E, Van Bourgondien ME, Wellman GJ, Love SR. Childhood Autism Rating Scale, Second Edition. Western Psychological Services; 2010.

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

Harrison PL, Oakland T. Adaptive Behavior Assessment System, Third Edition (ABAS-3). Western Psychological Services; 2015.

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.

Mokkink LB et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties. *Qual Life Res*. 2010;19(4):539–549.

Chan AW et al. SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. *Ann Intern Med*. 2013;158(3):200–207.

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

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

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