Psychometric foundations for cross-developmental adaptive assessment systems

Adaptive assessment systems spanning developmental stages require fundamentally different psychometric approaches than single-population tests. The evidence strongly supports **age-differentiated calibration strategies**, with 2PL IRT models for school-age children, Rasch for early childhood samples, and 3PL reserved for large-scale adult multiple-choice applications. Construct dimensionality demonstrably increases with age—what functions as a unified construct in early childhood differentiates into multidimensional structure by adolescence. Successful cross-developmental systems like the NIH Toolbox and MAP Growth achieve this through vertical scaling with common anchor items, age-appropriate stopping rules, and item banks spanning **5+ standard deviations** of difficulty. For implementation, the research consensus points to EAP estimation for interim scores transitioning to MLE/MLEF for final scoring, SE-based variable-length stopping rules (**SE ≤ 0.30** for clinical applications), and separate calibration with linking rather than concurrent calibration across wide age ranges.

IRT parameter calibration must adapt to developmental stage

Item Response Theory parameters behave differently across developmental populations, requiring calibration strategies tailored to each age range. The **discrimination parameter (a)** shows greater heterogeneity in early childhood due to developmental discontinuities, while **difficulty parameters (b)** must span narrower ranges for young children whose abilities cluster more tightly. The **guessing parameter (c)** becomes relevant primarily for multiple-choice formats with older children and adults.

For children aged 3-6, Rasch/1PL models provide superior performance with the smaller samples typically available, achieving stable estimates with **N = 100-200** examinees. Recent research on the WPPSI-IV with children around age 7 confirms that **2PL models consistently outperform 3PL and 4PL alternatives** for childhood cognitive assessment. Adolescent assessment benefits from 2PL with moderate discrimination variation, while adult standardized testing with large samples (N ≥ 2,000) supports full 3PL implementation. The emerging 4PL model, which adds an upper asymptote parameter for detecting careless responding, shows promise for children's cognitive tests where inattention creates characteristic response patterns. BVS Saúde

Sample size requirements scale dramatically with model complexity. Assessment Systems Stable 2PL parameter recovery requires **N ≥ 500-1,000**, while 3PL difficulty parameters only stabilize at **N ≥ 2,000**. For developmental assessments where sample constraints are common, Bayesian estimation with hierarchical priors offers an alternative path to stable parameters with smaller samples. Studies demonstrate that parameter recovery correlations exceed **r = 0.95** for both discrimination and difficulty when these minimums are met.

Differential Item Functioning demands systematic detection and handling

Age-related DIF occurs when items function differently across developmental groups at equivalent ability levels, threatening the validity of cross-age comparisons. Detection requires multiple complementary methods applied systematically across the item bank.

Mantel-Haenszel procedures provide efficient initial screening, detecting uniform DIF with samples as small as **200 per group**. The method yields a practical effect size (Δ-MH) with established classification thresholds: items exceeding $|\Delta\text{-MH}| > 1.5$ with statistical significance warrant removal or revision. **Logistic regression** extends detection to non-uniform DIF where the magnitude of group differences varies across ability levels, using a three-step model comparison approach. For large samples ($N \geq 500$-1,000 per group), **IRT-based methods** including likelihood ratio tests and Lord's χ^2 provide parameter-level insight into specifically which item characteristics differ between groups.

Handling strategies range from item removal to differential scoring. Pure removal simplifies scoring but risks content coverage gaps. More sophisticated approaches include **freeing parameters for DIF items during scoring** so that group-specific values are used, or implementing stratified item banks where flagged items are excluded from cross-age comparisons while retained for within-age assessment. Research on the EQ-5D-5L demonstrates that older adults show meaningful DIF on anxiety/depression items, being less likely to report problems at equivalent underlying health levels—illustrating how age can introduce systematic measurement bias requiring active management.

Separate item banks versus linking represents a fundamental design decision

The choice between maintaining separate age-specific item banks versus linking across ages through equating methods represents a central architectural decision with cascading implications for score interpretation.

Separate item banks optimize content appropriateness and avoid DIF complications, but preclude cross-age score comparisons and growth monitoring—a critical limitation for developmental assessment. **Linking approaches** enable growth tracking and more efficient item pool utilization but require construct equivalence verification and are vulnerable to DIF-induced validity threats.

Vertical scaling using common anchor items provides the most robust linking approach for developmental adaptive systems. Requirements include **20-25% anchor items** spanning the difficulty range between adjacent levels, with anchor items verified as DIF-free. Two primary calibration strategies exist: **concurrent calibration** estimates all parameters simultaneously (preferred when constructs remain stable), while **grade-by-grade calibration** with subsequent transformation proves more robust when constructs shift across development. Research by Wang and Jiao confirms that construct equivalence across grades must be empirically verified rather than assumed.

When construct meaning genuinely changes across development—termed **construct shift**—multidimensional vertical scaling using projection methods provides a valid alternative to standard unidimensional approaches.

This accommodates situations where, for example, early reading emphasizes decoding while later reading emphasizes comprehension, yet both require placement on a common developmental scale.

Construct dimensionality increases systematically across childhood

The **age differentiation hypothesis** receives consistent support across cognitive and developmental domains: abilities that function as relatively unified constructs in early childhood become increasingly differentiated with age. Tucker-Drob's analysis of 6,273 individuals ages 4-101 found that at lower ability levels, resource constraints lead multiple cognitive behaviors to covary strongly, while higher ability permits more differentiated cognitive profiles.

Swedish WPPSI-R standardization data for ages 3-7 demonstrates this pattern concretely—factor correlations between cognitive constructs decrease systematically as children develop. The ASQ-3 developmental screening tool shows multidimensional structure with five domains (communication, gross motor, fine motor, problem solving, personal-social), and studies confirm that multidimensional models provide superior fit across most age intervals from 6-60 months compared to unidimensional alternatives.

For adaptive assessment design, this implies that **unidimensional scoring may be appropriate for early childhood** while **multidimensional models become necessary for school-age assessment**. Item banks should accommodate both scoring frameworks, with bifactor structures—a general factor plus domain-specific factors—providing particular utility for generating both global and domain scores from the same item responses.

Floor and ceiling effects require proactive item bank design

Developmental assessment inherently risks floor effects in younger/lower-ability populations and ceiling effects as children mature beyond a test's original intended range. Standard detection thresholds identify concern when **>15% of scores fall at extremes** and severe problems when this exceeds 30%.

PROMIS research establishes the benchmark for adaptive solutions: the PROMIS Physical Function CAT covers **5.7 standard deviations** compared to legacy measures covering only 2.3 SDs, achieving **0% floor and 0% ceiling effects** versus 1-5% rates for fixed-form alternatives. This expanded range reduces required sample sizes by 25-50% for studies targeting floor populations.

The key implementation strategies include:

- Dedicated item development targeting extreme difficulty levels rather than just the middle range
- Graduated entry points based on developmental level or prior performance information
- Age-based item pool partitioning so initial items draw from developmentally appropriate difficulty ranges
- Early termination rules for healthy populations showing ceiling responses (PROMIS pediatric CATs terminate if first two items receive "healthiest" responses)

Starting and stopping rules require developmental customization

CAT algorithms must adapt their starting points and termination criteria to developmental populations for optimal efficiency and validity.

Starting rules for developmental populations should incorporate age-based priors rather than defaulting to population-mean ability estimates. Brief routing tests can establish initial ability estimates before the main CAT begins, particularly valuable for heterogeneous populations. The K-CAT system for child psychopathology screening exemplifies sophisticated starting approaches, tailoring initial question selection to both informant type and child age, with certain content areas (e.g., sexuality-related items) excluded below age 12. PubMed Central

Stopping rules for developmental CAT strongly favor variable-length approaches using standard error thresholds over fixed-length designs. PROMIS pediatric protocols use:

- Minimum: 4 items
- Maximum: 12 items
- SE threshold: 0.30 (theta metric) or 3.0 (T-score metric)
- Early termination for ceiling responses on initial items

Variable-length CATs achieve **~91% item reduction** while maintaining precision equivalent to full-length administration—critical for reducing fatigue effects in young children. The Predicted Standard Error Reduction (PSER) stopping criterion proves particularly useful for item banks with non-uniform information distribution, terminating when remaining items cannot meaningfully improve precision.

For item selection algorithms, **Kullback-Leibler Information** outperforms Fisher Information during early CAT stages, especially for extreme ability levels common in developmental populations. A-stratification with b-blocking reserves high-discrimination items for later stages when ability estimates have stabilized.

Ability estimation methods trade off bias against robustness

The choice between Bayesian and maximum likelihood estimation approaches has important implications for developmental CAT, where extreme ability levels and short test lengths are common.

Maximum Likelihood Estimation (MLE) produces unbiased estimates but cannot handle all-correct or all-incorrect response patterns, which occur frequently in early CAT stages and with developmental populations showing floor/ceiling performance. nih **MLE with Fences (MLEF)** addresses this limitation by incorporating

imaginary "fence" items at scale boundaries, producing estimates for all response patterns while avoiding Bayesian shrinkage toward the prior mean. (nih)

Expected A Posteriori (EAP) and **Maximum A Posteriori (MAP)** Bayesian methods always produce estimates but systematically pull scores toward the prior center—a bias that persists even after 30 items for examinees far from the prior mean. (nih) This shrinkage effect particularly threatens accuracy for developmentally delayed or advanced children whose true abilities lie in the distribution tails.

The practical consensus recommends a **hybrid approach**: EAP for interim estimation during early CAT stages when response pattern limitations are most problematic, transitioning to MLE or MLEF after 5-10 items when sufficient response variation enables likelihood-based estimation. For final scoring, MLEF offers the best combination of robustness and unbiasedness. (nih)

Validation requires multiple evidence sources applied systematically

CAT validation demands convergent evidence from content analysis, criterion relationships, and simulation-based precision evaluation. No single validation approach suffices for systems spanning developmental ranges.

Content validity requires expert review ensuring item representativeness across the assessed construct, with uniform distribution across measurement domains. **Criterion validity** typically correlates CAT estimates with established measures, with acceptable correlations ranging from r = 0.54-0.92 depending on stopping rules and criterion measure quality. **Simulation-based validation** using Monte Carlo methods with generated simulees tests algorithm performance under controlled conditions before live administration, with minimum recommendations of 1,000 simulees for adequate precision.

Key validation benchmarks from the literature include:

- Correlation between CAT theta and full item bank theta: **r > 0.88**
- Marginal reliability: **> 0.85-0.90**
- Test-retest reliability: CAT-Depression Inventory achieves **r = 0.92**, exceeding fixed-length PHQ-9 at r = 0.84
- Convergent validity with legacy instruments: PROMIS CATs demonstrate **r > 0.70** with SF-36 domains

Convergent and discriminant validity comparisons between adaptive and fixed-form assessments consistently favor CAT, particularly for measurement precision at ability extremes and floor/ceiling effect reduction.

Educational assessment shows mature CAT implementations with strong psychometric support

Reading and mathematics adaptive testing across K-12 have achieved operational maturity with robust validation evidence.

Reading CAT implementations face unique challenges from passage-based item structures requiring testlet management. Stanford's ROAR-CAT achieves **40% efficiency improvement**—75 CAT items match the reliability of 125 random items—with correlations of r = 0.89 against oral reading assessments. PubMed Renaissance's STAR Reading uses a 25-item adaptive test from 1,000+ vocabulary items, achieving split-half and test-retest reliability comparable to longer fixed forms, with correlations of **r = 0.91** against the Suffolk Reading Scale. Renaissance

Mathematics CAT systems span the developmental range from kindergarten through adult numeracy. MAP Growth provides grade-independent RIT scaling enabling growth tracking across years. NWEA ALEKS uniquely applies Knowledge Space Theory rather than IRT, with research showing students at 75% mastery are 2.7x more likely to achieve state assessment proficiency. McGraw Hill FastBridge aMath reduces testing time by 50-95% while matching 100-item fixed-form accuracy. Illuminate Education

The National Center on Intensive Intervention (NCII) provides independent validation ratings, with i-Ready Diagnostic achieving **"Convincing Evidence"** across all technical standards—the highest available rating. PR Newswire STAR assessments similarly receive highest NCII ratings for both screening and progress monitoring applications. Issuu

---

Cognitive and clinical CAT achieves substantial efficiency gains

Adaptive neuropsychological and clinical screening instruments demonstrate that CAT principles transfer successfully beyond educational achievement measurement.

The **NIH Toolbox Cognition Battery** spans ages 3-85, assessing six cognitive subdomains through seven computer-based measures. Adult validation (n = 268) shows test-retest reliability of **r = 0.86-0.92** for composite scores and convergent validity of **r = 0.78-0.90** with gold standard neuropsychological measures. The fluid cognition composite shows the expected strong negative correlation with age (r = -0.68), validating sensitivity to developmental and aging-related cognitive changes.

PROMIS CAT applies IRT-based adaptive testing to patient-reported outcomes across pediatric and adult populations. Pediatric PROMIS validation with 331 youth ages 8-17 shows high correlations between CAT and fixed short forms (**r = 0.79-0.92**) with average administration of only 4.7 items. CAT achieves wider accurate score ranges than short forms, with optimized stopping rules for populations with high proportions reporting best health.

CAT-MH provides comprehensive mental health screening across depression, anxiety, mania, substance use, PTSD, and other domains using multidimensional IRT. The K-CAT pediatric adaptation screens for seven

domains (depression, anxiety, mania, ADHD, conduct disorder, ODD, suicidality) PubMed Central in median **7.56 minutes** for child report and 5.03 minutes for parent report—compared to 125 minutes for the K-SADS-PL structured interview. Combined parent-child report achieves AUCs of **0.83-0.92** across diagnostic categories, with suicidal ideation detection reaching AUC = 0.996.

Implementation best practices span item bank construction through operational monitoring

Building cross-developmental adaptive systems requires attention to item bank architecture, calibration design, and ongoing quality monitoring.

Item bank construction should target pool sizes of approximately 12x the intended fixed-length equivalent—a 30-item CAT requires roughly 360 calibrated items. Item writing must uniformly cover the wide difficulty range needed for developmental spanning rather than concentrating on middle difficulty. "Enemies lists" prevent cluing between items, and field testing should embed within operational CAT for continuous bank development.

Calibration study design for developmental ranges requires adequate representation at each age level with sufficient overlap for linking. Common items between adjacent levels should comprise 20-25% of each level, spanning the difficulty range and verified as DIF-free. The decision between concurrent and separate-with-linking calibration depends on construct stability verification.

Operational monitoring includes:

- Regular DIF analysis across demographic groups including age Intensiveintervention
- Parameter drift detection with expert content review
- Item exposure monitoring with algorithm constraints
- Multiple item pools (primary + reserve) for security contingencies
- Simulation-based comparability verification when algorithms change

Score reporting should include confidence intervals reflecting conditional SEM, support both normative and criterion-referenced interpretations where appropriate, and present growth trajectories using vertically scaled scores alongside age-referenced interpretations.

Conclusion: Key implementation priorities

Cross-developmental adaptive assessment represents psychometric best practice when properly implemented. The evidence supports several clear priorities:

Model selection by development: Rasch for early childhood (ages 3-6) with small samples; 2PL for school-age children and adolescents with adequate samples ($N \geq 500$); 3PL reserved for adult standardized testing with

large samples ($N \geq 2{,}000$).

Dimensionality evolution: Design item banks supporting unidimensional scoring for young children transitioning to multidimensional scoring for older populations. Bifactor structures provide flexibility for both global and domain-specific interpretations.

Vertical scaling architecture: Use common anchor items (20-25% overlap) with DIF verification, separate-with-linking calibration when construct equivalence cannot be assumed, and projection methods when genuine construct shift occurs.

Adaptive algorithm customization: Age-based priors for starting estimates, SE-based variable-length stopping (target $SE \leq 0.30$), early termination for ceiling/floor responses, and Kullback-Leibler information for item selection in developmental populations.

Estimation strategy: EAP for interim estimates during early CAT stages, transitioning to MLEF or MLE for final scoring to avoid Bayesian shrinkage at ability extremes.

Major operational systems including MAP Growth, STAR, NIH Toolbox, PROMIS, and CAT-MH demonstrate these principles achieve **50-90% testing time reduction** while maintaining or exceeding fixed-form reliability and validity. The psychometric foundations for cross-developmental adaptive assessment are well-established—successful implementation requires systematic attention to the developmental considerations that pervade every design decision from item calibration through score interpretation.