WORKING WITH THE THRIVE BY FIVE INDEX 2024:
EXPLORATIONS OF EARLY LEARNING SYSTEMS IN SOUTH AFRICA

ON-TRACK IN EARLY CHILDHOOD: MACHINE LEARNING PREDICTION OF EARLY LEARNING STATUS IN THE CONTEXT OF SOCIOECONOMIC AND DEVELOPMENTAL RISK

MICHELLE LEAL

WORKING WITH THE THRIVE BY FIVE INDEX 2024:
EXPLORATIONS OF EARLY LEARNING SYSTEMS IN SOUTH AFRICA

ON-TRACK IN EARLY CHILDHOOD: MACHINE LEARNING PREDICTION OF EARLY LEARNING STATUS IN THE CONTEXT OF SOCIOECONOMIC AND DEVELOPMENTAL RISK

MICHELLE LEAL

SAMRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, School of Clinical Medicine, University of the Witwatersrand, Johannesburg, South Africa.

Abstract

Background: Early learning shortfalls remain widespread in low- and middle-income countries, but evidence is limited on which indicators best predict on-track early learning among children exposed to adversity. The objective was to quantify predictive signal for early learning status in a nationally sampled South African cohort.

Methods: Cross-sectional data from the 2024 Thrive by Five Index were analysed for children aged 50–59 months exposed to socioeconomic or growth-related adversity (n=2278). On-track early learning status was derived from the Early Learning Outcomes Measure (ELOM 4&5). A Mixture-of-Experts neural network with grouped cross-validation was benchmarked against histogram-based gradient boosting and elastic-net logistic regression. Domain and source ablation analyses assessed contributions of child, caregiver, and programme predictors.

Results: Discrimination was modest. Histogram-based gradient boosting achieved the highest performance (AUC 0.66, SD 0.02), followed by the Mixture-of-Experts model (AUC 0.62, SD 0.02) and elastic-net logistic regression (AUC 0.60, SD 0.03). Predictive signal was concentrated in child and caregiver indicators, particularly age, home language, caregiver-reported developmental status, and household learning resources. Programme predictors contributed limited incremental signal.

Interpretation: Within adversity-exposed populations, routinely collected child and caregiver indicators may support stratification for additional assessment or support. Programme-level targeting will require improved measurement of enacted pedagogical quality and external validation across settings.

Research in Context

Evidence before this study

Developmental shortfall remains widespread in low- and middle-income countries, with an estimated 250 million children younger than 5 years not reaching age-appropriate milestones [1,2]. Cognitive and language outcomes show strong gradients by socioeconomic conditions, caregiver stimulation, and linear growth [3,4]. Large-scale analyses have relied primarily on child and household survey measures, with limited linkage to programme-level indicators [3]. Evidence from early childhood education research indicates that structural programme characteristics show weak or inconsistent associations with child outcomes, whereas measures of process quality—particularly teacher–child interaction and instructional support—are more consistently associated with development [5,6]. Teacher-training trials further show that improvements in pedagogical practice can improve early learning outcomes when implementation intensity is adequate [7,8].

Added value of this study

Using a nationally representative South African cohort restricted to children exposed to socioeconomic or growth-related adversity, grouped cross-validated machine learning was used to quantify predictive contributions from child, caregiver, and programme domains. Predictive signal was concentrated in age, growth-related markers, and caregiver-linked resources, whereas programme descriptors contributed limited incremental information.

Implications of all available evidence

Within adversity-exposed populations, routinely collected child and caregiver indicators may support stratification for additional developmental assessment or support. Programme-level targeting will require improved measurement of pedagogical process quality and validation across settings.

Background

Developmental shortfall—defined as the failure to meet age-appropriate cognitive, language, socio-emotional, and motor milestones—represents a primary attrition of human capital in low- and middle-income countries (LMICs). Globally, 250 million children under 5 years (43%) fail to reach developmental potential [1,2]. These deficits arise from the co-occurrence of household poverty, nutritional deprivation, and inadequate early learning opportunities. From a life-course perspective, such early-life exposures consolidate into persistent disadvantage; altering developmental trajectories such that by age 10, 57% of children in LMICs lack basic reading comprehension, compared with 9% in high-income settings [9]. These trajectories are predictive of fewer years of completed schooling and diminished adult earnings [1,10].

Sub-Saharan Africa remains the region of highest risk, where stunting affects nearly one in three children under five [11]. Pre-pandemic regional learning poverty stood at 86%; pandemic-related school closures have since increased this burden to nearly 70% across LMICs [9].Access to quality early childhood care and education (ECCE) is a primary bottleneck. Only 25% of children aged 3–4 years in sub-Saharan Africa are enrolled in early learning programmes, versus 90% in high-income settings [3,12].

South Africa demonstrates a disconnect between service coverage and developmental outcomes. In 2021, 81% of Grade 4 learners could not read for meaning, exceeding the LMIC average of 57% [13]. Educational efficiency is similarly low; children accrue only 5.6 learning-adjusted years of schooling (LAYS), significantly below outcomes in Kenya (8.5) and Ghana (6.0) [14,15]. Data from the Thrive by Five Index indicate that fewer than half of children aged 4–5 years are simultaneously on track for both growth and early learning outcomes [16]. Stunting prevalence remains stagnant at 20%, lower than Rwanda (33%) but higher than Kenya (18%) and Ghana (17%) [11].

Despite sustained policy focus, early childhood interventions often remain broad and insufficiently differentiated, thereby failing to support the most vulnerable populations [1,17,18]. Operational definitions and thresholds for "resilience" are inconsistent and rarely implemented at population level [19,20]. National averages obscure socioeconomic gradients, and integrated child–caregiver–programme profiles that identify on-track early learning within adversity are scarce in LMIC contexts [2,3]. This constrains SDG 4.2.1 baseline setting and targeted service expansion [21,22]. The objective was to quantify the distribution of explanatory factors for early learning status across ecological levels within an adversity-defined national cohort. Using a nationally representative survey of South African children aged 50–59 months, the analysis examined on-track early learning within children exposed to socioeconomic or growth-related risk. The analysis evaluated the relative contribution of child, caregiver, and programme domains to predictive utility, aiming to generate decision-relevant profiles for targeted early childhood support.

Methods

Study design

The 2024 Thrive by Five Index utilised a national, cross-sectional design anchored in the Nurturing Care Framework for early childhood development in South Africa [18]. A stratified, three-stage probability design (ward → ELP → child) was implemented. Primary sampling units (PSUs) comprised wards, allocated to ensure each of the nine provinces received a minimum of 35 PSUs, with the remainder distributed proportional to the population of two-year-olds in the 2022 National Census. Selection within strata utilised probability-proportional-to-size (PPS) based on Grade 3 learner counts. Fieldwork in 2024 yielded 1,388 ELPs and 5,001 children, exceeding design targets after the exclusion of assessments contravening Early Learning Outcomes Measure (ELOM 4&5) quality protocols [23–25].

Three survey weights—child, ELP, and primary caregiver—accounted for the inverse of stage-specific selection probabilities and incorporated adjustments for on-site child replacement and ELP additions. Population-level inference was recovered by specifying design information, including stratification and clustering, for all weighted estimates. The multistage selection probability was calculated as the product of: (i) PSU selection with probability proportional to size within province-by-quintile strata; (ii) ELP selection within PSUs based on a comprehensive listing; and (iii) within-ELP child selection from a sex-stratified random rank list (target: two girls and two boys per ELP), with systematic recording of on-site replacements [23,24].

In instances where fewer than 12 children were assessed within the initial three ELPs of a PSU, an additional ELP was selected within the same PSU or stratum to reach the child target, with weights adjusted accordingly. All descriptive and inferential estimates were design-based, specifying stratification and clustering at both ward and early learning programme levels.

Setting and Population

Data collection occurred within early learning programmes across diverse urban and rural settings in all nine South African provinces during 2024. The total analytic sample included 5,001 children, 1,388 programmes and principals, and 1,223 practitioners. Linked caregiver interviews were completed for 3,841 children, representing 77% of the assessed sample [23,24].

In instances where fewer than 12 children were assessed within the initial three ELPs of a PSU, an additional ELP was selected within the same PSU or stratum to reach the child target, with weights adjusted accordingly. All descriptive and inferential estimates were design-based, specifying stratification and clustering at both ward and early learning programme levels.

Setting and Population

Data collection occurred within early learning programmes across diverse urban and rural settings in all nine South African provinces during 2024. The total analytic sample included 5,001 children, 1,388 programmes and principals, and 1,223 practitioners. Linked caregiver interviews were completed for 3,841 children, representing 77% of the assessed sample [23,24].

Eligibility criteria required children to be aged 50–59 months and present at the facility on the day of the visit. Children with severe sensory or mobility impairments were excluded. Documentation of caregiver consent and child assent was a prerequisite for participation [23].

To evaluate developmental on-track status conditional on structural disadvantage, the dataset was strictly restricted to children meeting established early-life adversity criteria. Following the standardisation of variables, 2723 children were excluded for falling below the adversity threshold. The primary on-track early learning outcome was fully observed across the remaining cohort, yielding a final restricted analytic sample of N=2278 (Figure 1).

Demographic and Socioeconomic Variables

Predictors were derived from six Thrive by Five instruments: the child sampling form, the practitioner-rated Social Emotional Functioning (SEF) scale, the primary caregiver interview, paired practitioner and principal interviews, the facility observation checklist, and the Learning Programme Quality Assessment (LPQA) [24,26]. Direct child assessment (ELOM 4&5) variables were excluded from the primary classification models to prevent circularity, as the developmental on-track status outcome was derived from these performance categories. Household asset deprivation was used only to define the adversity-restricted analytic sample and was not included in the outcome definition. These features were utilised exclusively in separate pipeline-verification models.

Child-level determinants. Province was harmonised from survey identifiers and sex was coded as a binary variable from assessment records. Nutritional status was defined using height-for-age z-scores (HAZ), summarised as a binary indicator of stunting (< -2 SD) versus not stunted (≥ -2 SD) [27]. Perinatal indicators were linked from the caregiver Road-to-Health Booklet (RtHB), including birthweight (low-birthweight flag, <2,500 g), gestational age (preterm flag, <37 weeks), and a binary indicator for availability of the RtHB at the time of the interview. Multiplicity of home languages was derived from child-level checkboxes and coded as an indicator for two or more home languages.

Household and caregiver context. The caregiver module provided data on educational attainment, employment status, internet access, home learning resources, perceived social support, dwelling type, social grant receipt, and household assets. Education was recorded in categories from primary schooling or less to degree-level qualifications. Employment status included categories for the employed, the unemployed, those not in the labour force, and students. Internet access was distinguished by connection type, and home learning stimulation was measured by the number of children's books available in the home. Family social support was recorded using a Likert response scale. Dwelling formality was derived from structural housing types, and child social grant receipt was recorded as a binary indicator. Household assets were analysed as a summative index of binary asset indicators.

Early learning programme context. Structural features included setting and infrastructure indicators from the facility observation (e.g., location mobility, building type, kitchen availability, food refrigeration, and security measures). Water, sanitation, and hygiene (WASH) status was recorded as source categories and summarised using Joint Monitoring Programme (JMP) proxies for improved versus unimproved services.

Process quality was derived from the LPQA classroom observation, encompassing ordered three-level ratings (Inadequate, Basic, Good) for the overall score and five specific domains: materials and equipment, planning and assessment, the learning programme, teaching strategies, and relationships and interactions. LPQA descriptors included class size, the presence of assistants, the main language of instruction, and the proportion of time the main language was used during the observation.

Educator perspectives. Centre-level information from paired principal and practitioner interviews included educator qualifications, job roles, curriculum availability, and perceived support. Variables further captured parental-engagement practices, programme finance sources, and employment conditions, including contract status, salary bands, and leave benefits. Attitudinal items regarding pedagogy and growth-mindset beliefs were recorded as ordered four-point factors. This multi-layered predictor set was utilised to evaluate the interplay between risk factors and protective resources in relation to early learning status.

Early Learning On-track Status

Early learning on-track status was measured using the Early Learning Outcomes Measure (ELOM 4&5), a South African direct-assessment tool standardised across 11 official languages. The instrument evaluates five developmental domains: gross motor, fine-motor/visual-motor integration, emergent literacy, emergent numeracy, and executive functioning. Performance was analysed using composite scores on a 0–100 scale and norm-based categories ("On Track", "Falling Behind", and "Falling Far Behind").

Primary Scale of Intelligence—Fourth Edition (WPPSI-IV) Full-Scale IQ is moderate ($r \approx 0{\cdot}64$) [28]. Internal consistency for the SEF total in the 2024 sample was $\alpha = 0{\cdot}92$ (Cronbach).

Outcome Measures

The primary outcome was a binary "on-track" indicator, where 1 represented the "On Track" composite performance category and 0 otherwise. Social-emotional status was recorded via a 28-item practitioner-rated Social-Emotional Functioning (SEF) scale, encompassing relations with peers and adults, emotional regulation, prosocial behaviour, and task focus. These data were utilised for divergent-validity checks of the ELOM 4&5 cognitive and motor domains. To ensure model integrity and prevent informational leakage, all continuous ELOM domain and composite scores were excluded from the predictor set as these were used to derive the target labels. Similarly, the SEF total and sub-scores were excluded to prevent source bias, as these practitioner-rated measures may reflect subjective perceptions that co-vary with the assessment environment.

Definition of on-track early learning status within adversity

Consistent with developmental-systems theory [19], on-track status was operationalised as achievement of an "On Track" early learning classification among children exposed to significant ecological adversity. Children were classified as on track for early learning within adversity if they met the primary early learning outcome (binary on-track flag = 1) while simultaneously meeting at least one of two adversity criteria: (i) residence in a household within the lowest national weighted asset quintile; or (ii) evidence of growth risk, defined as a height-for-age z-score (HAZ) below -1 SD.

The −1 SD threshold was used to capture early growth faltering associated with developmental vulnerability, rather than restricting adversity to moderate or severe stunting (HAZ < −2 SD). Evidence from LMIC cohorts indicates that cognitive and developmental outcomes decline progressively across the HAZ distribution and are not confined to the clinical stunting threshold [4,29,30]. This dual criterion identified positive adaptation in the face of concurrent socioeconomic and developmental precariousness. While this definition served as the basis for design-based descriptive analyses of children who were on track for early learning within the adversity-exposed group, the predictive machine-learning models were fitted specifically for the early learning outcome within that same adversity-defined sample.

Statistical Analyses

Analyses of descriptive characteristics used survey-weighted methods in R (version 4.4.3) to account for the stratified, clustered sampling design, with module-specific weights applied according to data source (child, caregiver, and programme-level modules). The household asset distribution (including the lowest asset quintile threshold) was estimated using caregiver weights. Variance estimates were obtained by Taylor series linearisation, and singleton primary sampling units were handled using the survey package's adjustment option (survey.lonely.psu = "adjust"). Results are reported as unweighted frequencies and survey-weighted proportions or estimates.

Machine-learning analyses were conducted separately in Python (version 3.12.3) and, in the primary analyses, models were trained without sampling weights because the estimand was individual-level predictive performance (discrimination and calibration) within the adversity-defined analytic cohort, rather than estimation of a population-average parameter [31–33]. Predictive performance was assessed using out-of-fold predictions from grouped cross-validation with grouping at the ECD-centre level (id_ecd) to reduce within-centre information leakage. Calibration and discrimination were evaluated using pooled out-of-fold predictions from the cross-validation procedure. Discrimination was quantified using the area under the receiver operating characteristic curve (AUC) and average precision (AP), and overall probabilistic accuracy was summarised using log loss. Calibration performance was assessed using the Brier score together with calibration intercept and calibration slope. Calibration intercept and slope were estimated using logistic recalibration regression of the observed outcome on the logit of predicted probabilities, where an intercept of 0 and slope of 1 indicate perfect calibration. Calibration curves were constructed by grouping predicted probabilities into deciles and plotting observed outcome frequencies against mean predicted probabilities within each bin.

To assess the impact of this modelling choice under a complex survey sampling frame, a sensitivity analysis compared the primary unweighted specification with weighted evaluation metrics and weighted-loss training specifications. Results of these analyses are reported in Supplementary Table S4 [34].

Machine Learning Architecture

Predictive modelling was restricted to the adversity-defined analytic sample to examine predictors of on-track early learning status within children exposed to socioeconomic or growth-related risk. The primary classification target was the binary on-track early learning status outcome. Classification utilised a sparse, top-k Mixture of Experts (MoE) neural network [35], comprising three multi-layer perceptron experts and a gating network. Expert routing was implemented via sparse top-k gating (k=2), in which only the two highest-probability experts were activated for each observation [36].

For the Mixture-of-Experts model, routing statistics including expert utilisation frequencies, routing entropy, and gating confidence were extracted from the trained model across cross-validation folds.

The three-expert configuration provided sufficient model capacity while mitigating subnetwork under-utilisation relative to the analytic sample size [37]. Model components were trained jointly via an AdamW optimiser (decoupled weight decay 10^{-4}) with an auxiliary load-balancing loss to maintain expert diversity. Each expert consisted of a two-layer multilayer perceptron (128 and 64 hidden units) with ReLU activation functions and dropout (0.10).

Data preparation and Covariate Selection

Predictors were derived from six survey modules, with categorical variables coded as dummy indicators. Sentinel codes were recoded to null prior to analysis. Predictors were derived from six survey modules, with categorical variables represented via one-hot encoding. Sentinel codes were recoded to null prior to analysis. Missing observations were handled using iterative regression-based imputation (IterativeImputer, scikit-learn) [40], modelling each missing value as a conditional function of the remaining feature set to preserve the multivariate covariance structure and reduce attenuation bias in feature importance estimates. To prevent information leakage, imputation was performed within each cross-validation training fold: the imputer was fit on the training partition and then applied to the corresponding validation fold.

A staged feature audit was implemented to identify predictors that could introduce information leakage or structural redundancy. As a diagnostic safeguard, each predictor was screened individually for unusually high predictive power using single-variable AUC; predictors exceeding AUC ≥ 0.80 (or ≥ 0.70 for child-level indicators) were reviewed and removed if identified as outcome-defining or otherwise indicative of information leakage. This step served as a data-quality audit rather than a supervised feature-selection procedure used to optimise model performance. Dimensionality reduction further included removal of near-duplicate numeric columns ($|r| > 0.98$) and ultra-sparse features (<1% observed).

To reduce the influence of predictors that directly capture the immediate testing context, assessor observations of child test-taking behaviour and practitioner-rated socio-emotional scores were excluded from the feature space. These variables reflect behaviours observed during the assessment itself and could therefore dominate model predictions without representing broader determinants of early learning status. Variables used in the derivation of the outcome label were similarly removed to prevent circularity and measurement leakage between predictors and the target.

Model Interpretability and Ablation

Predictor contributions were estimated via mean absolute SHAP (SHapley Additive exPlanations) values, aggregated across cross-validation folds to ensure global consistency [41]. Relative predictive utility across the six survey modules was isolated through systematic ablation. Differences in classification performance were recorded under both source-restricted and leave-one-source-out configurations to isolate the information signal provided by each instrument.

Results

Descriptive statistics

Sample Characteristics and Biological Factors

The mean age of the cohort was 54.8 months (SD 2.4), balanced by sex (50.9% female). Anthropometric assessment indicated a mean Height-for-Age z score (HAZ) of -1.1 (SD 0.9) and a stunting prevalence of 14.2% (Table 1). While the mean HAZ did not differ between groups (p = 0.869), children who were on track for early learning were less likely to be stunted compared to those who were not on track (10.9% vs 16.2%; p = 0.014). Children who were on track were also slightly older (mean 55.5 vs 54.4 months) and more likely to be female (55.9% vs 48.0%; p = 0.009). Linguistic diversity was low, with only 2.9% of households speaking more than one language, though this was more common among children who were on track (p = 0.043). Sensitivity analyses using alternative HAZ thresholds (-1.0, -1.5, and -2.0 SD) showed that stricter growth-risk cut-offs reduced the adversity subgroup size (45.6%, 34.0%, and 27.2% of the recoded sample, respectively) with only modest changes in the proportion classified as on track for early learning within the adversity subgroup (38.3%, 36.4%, and 34.9%; Supplementary Table S3).

Household Environment and Early Learning Status

Caregiver and household characteristics revealed distinct socioeconomic gradients between groups. Households of children who were on track for early learning had significantly higher asset scores (mean 5.4 vs 5.1; $p = 0.035$) and caregivers were more likely to be employed (45.7% vs 38.5%; $p = 0.031$). Access to early learning materials differed between groups; over a third (35.5%) of the group not on track for early learning had no children's books in the home, compared to 24.8% of the group on track for early learning ($p = 0.002$). Furthermore, 98.5% of children who were on track for early learning possessed a Road to Health Booklet, a proxy for healthcare engagement, compared to 96.1% of those not on track ($p = 0.032$).

Programme Context and Institutional Quality

Most children were captured in ECD centres (73.1%) rather than home-based or community settings. While programme registration and province did not significantly differ by early learning status, financial barriers were evident; children who were on track for early learning were more likely to attend high-fee programmes (7.9% vs 3.2%; $p = 0.004$). Structural and process quality measures—including mains electricity (77.3%), secured perimeters (96.0%), and the Total Learning Programme Quality Assessment (LPQA) scores—were consistent across both groups ($p > 0.05$). The institutional quality remained relatively homogenous across the sample, with the mean LPQA score at 34.3 (SD 13.0), and staff-child interaction scores not varying by on-track status (mean 7.8; $p = 0.859$).

Educator Human Capital and Pedagogical Beliefs

Practitioner characteristics were largely similar between groups. Approximately half of the practitioners had received training in the previous 12 months, and the majority (80.4%) demonstrated awareness of the National Curriculum Framework (NCF). Furthermore, over 90% of practitioners across both groups exhibited a "Growth Mindset." Practitioner remuneration was low, with 39.1% reporting monthly earnings below R1500. The distribution of salary categories did not differ by the early learning status of the children in their care.

Model Performance

Model performance for the full feature set is shown in Table 2. In unweighted cross-validation, the histogram-based gradient boosting (HGB) model had an AUC of 0.66 (SD 0.02), followed by the Mixture-of-Experts (MoE) model (AUC 0.62, SD 0.02) and elastic-net logistic regression (AUC 0.60, SD 0.03). Average precision values were 0.55 (SD 0.01) for HGB, 0.51 (SD 0.02) for MoE, and 0.49 (SD 0.04) for elastic-net logistic regression. Log loss values were 0.64 (SD 0.01) for HGB, 0.65 (SD 0.01) for MoE, and 1.07 (SD 0.10) for elastic-net logistic regression. Sensitivity analyses applying sampling weights to evaluation metrics and training loss produced similar discrimination and calibration estimates (Supplementary Table S4).
[Table 2]

Model Discrimination

Receiver operating characteristic and precision–recall curves derived from the cross-validation procedure are shown in Figure 2. Discrimination was highest for the histogram-based gradient boosting model (AUC 0.66; AP 0.55), followed by the Mixture-of-Experts model (AUC 0.62; AP 0.51) and elastic-net logistic regression (AUC 0.60; AP 0.49).

Model Calibration

Calibration of predicted probabilities for the histogram-based gradient boosting, Mixture-of-Experts, and elastic-net logistic regression models is shown in Figure 3. The histogram-based gradient boosting model had the lowest Brier score (Brier score 0.221; calibration intercept −0.074; calibration slope 0.661). Calibration was similar for the Mixture-of-Experts model (Brier score 0.229; calibration intercept −0.144; calibration slope 0.658). The elastic-net model showed substantially poorer calibration and probability resolution (Brier score 0.308; calibration intercept −0.349; calibration slope 0.126). Calibration slopes were below 1 for all models. Calibration curves showed wider dispersion in the highest predicted-probability bins, reflecting smaller numbers of observations in the highest predicted-risk groups.

Mixture-of-Experts Routing Behaviour

Routing statistics were examined to assess expert utilisation within the Mixture-of-Experts architecture. Across cross-validation folds, expert selection was distributed across all three experts, with mean top-1 routing frequencies of 0.45, 0.26, and 0.30 for Experts 1–3 respectively. Routing entropy was high (mean 1.10; normalised entropy 0.999), indicating diffuse routing across experts consistent with the sparse top-k (k=2) gating mechanism. The mean routing confidence was 0.35, indicating that multiple experts contributed to model predictions rather than a single dominant expert. Routing patterns were stable across cross-validation folds, with minimal variation in entropy and expert utilisation.

Covariate Contributions

Permutation importance estimates from the histogram-based gradient boosting model showed the largest contribution for child age (importance 0.055), followed by caregiver-reported home language (0.011). Additional important predictors included child sex recorded in the ELOM assessment (0.0048), caregiver-rated child developmental status relative to peers (0.0047), and caregiver-reported child sex (0.0035) (Figure 4). Additional predictors included the availability of handwashing facilities at the programme (0.0031), total programme staffing (0.0022), frequency of caregiver storytelling or reading to the child (0.0021), practitioner-reported programme quality support (0.0019), and the number of children's books in the home (0.0016). Remaining predictors demonstrated smaller contributions (<0.0015). SHAP (SHapley Additive exPlanations) analyses showed broadly similar ordering of the highest-contributing predictors across cross-validation folds (Supplementary Figure S1).

Domain-restricted Analyses

Domain-restricted analyses were estimated using the Mixture-of-Experts model (Supplementary Table S1). Relative to the full model (AUC 0.62), removal of caregiver predictors produced minimal change in discrimination (ΔAUC ≈ -0.003) and a small increase in log loss (Δlog loss $\approx +0.01$). Removal of child predictors reduced discrimination by approximately 0.02 AUC and increased log loss by approximately 0.02. Removal of predictors classified as "other" did not materially affect model performance. In contrast, removal of early learning programme (ELP) predictors produced a small increase in discrimination (ΔAUC $\approx +0.01$) and a slight reduction in log loss.

In single-domain ("domain-only") analyses, child predictors yielded the highest discrimination (AUC 0.63), followed by caregiver predictors (AUC 0.59) and ELP predictors (AUC 0.59).

Source-level Incremental Predictive Utility

Incremental predictive utility was estimated using the Mixture-of-Experts model (Supplementary Table S2). Relative to the full model (AUC 0.62), removal of caregiver predictors produced minimal change in discrimination (ΔAUC ≈ −0.003). Removal of facility predictors did not materially affect discrimination (ΔAUC ≈ 0.00). Removal of LPQA predictors reduced discrimination by approximately 0.01 AUC, whereas removal of principal predictors produced a similarly small reduction (ΔAUC ≈ −0.003). In contrast, removal of practitioner predictors resulted in a small increase in discrimination (ΔAUC ≈ +0.01).

In single-source ("source-only") analyses, caregiver predictors yielded the highest discrimination (AUC 0.59, SD 0.03), followed by LPQA predictors (AUC 0.59, SD 0.01) and principal predictors (AUC 0.59, SD 0.02). Facility-only and practitioner-only models demonstrated lower discrimination (AUC 0.54 and 0.52, respectively).

Discussion

In this nationally sampled cohort restricted to socioeconomic or growth-related adversity, discrimination for early learning on-track classification was modest. Histogram-based gradient boosting achieved the highest discrimination (AUC 0.66), followed by the Mixture-of-Experts model (AUC 0.62) and elastic-net logistic regression (AUC 0.60). Because the primary objective was to quantify the contribution of predictors across ecological levels, the Mixture-of-Experts architecture was retained for domain-structured analyses, where its modular structure permits systematic ablation of child, caregiver, and programme predictors. Performance of this magnitude is consistent with prediction tasks based on cross-sectional household and programme descriptors rather than repeated developmental assessments or direct measures of caregiving practice and pedagogy, which capture more proximal determinants of early learning [1]. Predictive information in this analysis was concentrated in child and caregiver domains, and model outputs are therefore more appropriately interpreted as tools for population stratification rather than individual classification [42]. The analysis quantifies explanatory power within the measured indicators and should not be interpreted as evidence of causal determinants of developmental status [43]. This is also consistent with the structure of nationally representative LMIC datasets, in which developmental risk analyses rely predominantly on child and household measures, with limited availability of linked programme-level process indicators at scale; analyses combining an adversity-restricted analytic frame with grouped cross-validation and domain-level attribution across ecological sources remain uncommon [3,10,21].

Within this overall pattern, age in months was the highest-ranked predictor despite restriction to a 50–59-month eligibility band. Relative-age effects within school-entry cohorts are well documented in early developmental assessments, particularly when categorical performance bands are used [44]. Differences of several months within a single age year are associated with lower executive function, language, and early numeracy among younger children in the same cohort, with larger gradients observed in socioeconomically disadvantaged populations [44]. Within the restricted age range used in this study, variation of several months may therefore influence classification around categorical on-track thresholds. These findings highlight the importance of accounting for age in months when applying categorical readiness classifications in high-risk populations.

After age, growth-related indicators remained prominent despite restriction to adversity and despite stunting not defining inclusion. Stunting prevalence was lower among children who were on track for early learning (10.9% vs 16.2%), whereas mean HAZ did not differ between groups. Linear growth is widely interpreted as a cumulative marker of early-life constraint that co-varies with developmental attainment, although mean differences may be attenuated in restricted

samples. Meta-analyses of LMIC cohorts indicate that a 1 SD increase in height-for-age is associated with approximately 0·20 SD higher cognitive scores, with stronger gradients observed where multiple risks co-occur [4]. These associations likely reflect shared antecedents—including nutrition, infection burden, caregiving environments, and socioeconomic conditions—rather than a single causal pathway operating at school entry [21,30] . Within an adversity-defined sample, the persistence of growth indicators among leading predictors is therefore consistent with cumulative early-life exposure rather than a contemporaneously modifiable determinant of developmental status.

Household-linked factors showed a similar pattern. Markers of stimulation and learning resources—particularly the presence of children's books—differentiated children who were on track from those who were not, with a substantial contrast in the proportion reporting no books in the home (24.8% vs 35.5%). Home learning environments and caregiver engagement are strongly patterned by socioeconomic status and are consistently associated with early cognitive and language outcomes across LMIC settings [2]. Pooled analyses indicate that children exposed to multiple home learning activities have developmental scores approximately 0.3–0.5 SD higher than those with minimal stimulation, with gradients persisting after adjustment for preschool attendance [45,46]. Global monitoring data further indicate that inequalities in access to early learning opportunities have narrowed only modestly over the past decade [45]. Within an adversity-defined sample, household-linked inputs therefore remain measurable sources of variation in early learning status.

The concentration of predictive signal in a small number of proximal indicators was also evident in the feature attribution analyses. After age in months, child sex and caregiver-reported developmental status ranked among the highest contributors to model predictions, indicating measurable signal from maturational timing and caregiver-reported functioning within the adversity-defined sample. Caregiver-report instruments have shown moderate agreement with direct developmental assessments and have been used to characterise early development in large-scale surveys, supporting their utility for population-level measurement when direct testing is not feasible [47]. However, such measures may also capture caregiver perception and reporting bias. Sex differences in early developmental outcomes have also been reported in multi-country analyses, with small but detectable average differences that persist after accounting for household and socioeconomic circumstances [48].

Beyond these leading predictors, contextual contributions were smaller but not absent. Programme staffing ranked above most other institutional indicators, which is consistent with evidence that structural indicators such as staff–child ratios show variable and generally modest associations with child developmental outcomes compared with measures of process quality

[49,5]. Household language and caregiver age were among the highest-ranked caregiver-linked variables, aligning with broader evidence that language environments and caregiver characteristics are associated with early developmental opportunity structures across LMIC settings [47]. Environmental proxies (for example water source and programme location) contributed little to model performance, which may reflect the limited resolution of coarse categorical measures for capturing distal structural conditions that influence child development through indirect pathways such as infection burden, caregiving environments, and nutrition [50]. Feature attribution was highly concentrated, with most remaining variables contributing <0.002 in permutation importance.

Taken together, these findings indicate that programme-domain predictors contributed limited incremental predictive signal within the measured feature set. Domain ablation showed that removal of early learning programme (ELP) variables produced a small increase in discrimination (AUC ≈ 0.62 to 0.64) and a slight reduction in log loss, whereas models trained using ELP predictors alone showed lower discrimination (AUC ≈ 0.59). Similar patterns were observed for facility-only and practitioner-only specifications, which demonstrated limited predictive performance relative to models including child and caregiver variables. LPQA predictors provided modest discrimination in isolation but did not materially improve performance when combined with stronger proximal predictors. In predictive modelling, such patterns often arise when predictors contain limited additional signal relative to variables already included in the model, or when measurement resolution is insufficient to capture meaningful variation across units [42]. Evidence from early childhood education research indicates that structural characteristics—such as infrastructure or staffing ratios—often show weak or inconsistent associations with child developmental outcomes, whereas measures of process quality, including teacher–child interaction and instructional support, demonstrate more consistent relationships with early learning [6].

This interpretation is consistent with evidence from intervention studies. Improvements in pedagogical practice can produce measurable gains in school readiness, whereas structural inputs alone are often insufficient. In the Quality Preschool for Ghana cluster-randomised trial, teacher training and coaching improved classroom practices and were associated with higher language and executive function scores, with sustained effects where subsequent environments supported consolidation [7]. Early childhood teacher-training interventions in other LMIC settings have similarly improved classroom quality and child behaviour when practice-level change and implementation intensity were adequate [8,17]. Within the TB5 feature set, programme variables did not materially improve predictive discrimination.

Two explanations are plausible. First, although the LPQA instrument was designed to capture classroom process quality—including domains such as teaching strategies, planning and assessment, and relationships and interactions—the available indicators may not capture enacted pedagogy at sufficient resolution to contribute strong predictive signal in this modelling context, consistent with evidence that process quality measures are more strongly associated with early learning outcomes than structural programme indicators [5,51]. Second, although the present analysis cannot directly evaluate sorting between households and programmes, programme-level effects may be partially absorbed by child and household characteristics if children experiencing greater socioeconomic constraint are more likely to attend lower-resourced programmes, a pattern widely documented in early childhood systems globally [17,21].

Model comparison points in the same direction. Histogram-based gradient boosting showed higher discrimination than elastic-net regression within the available predictor set. Absolute performance remained modest, and precision–recall profiles indicated substantial trade-offs between sensitivity and positive predictive value at thresholds suitable for individual classification. Under these conditions, model outputs are more appropriately interpreted as tools for stratification to prioritise additional assessment or targeted support rather than for binary screening [42]. AUC alone is insufficient for decision-making; useful implementation requires adequate calibration, thresholds aligned to programme capacity, and evaluation of downstream benefits and costs [42,52].

Calibration is central to that distinction. In the present analysis, probability estimates were well calibrated in the central risk range with mild over-prediction at the highest decile, supporting use for stratification while reinforcing that individual classification would require further implementation evaluation. The out-of-fold Brier score for the primary model was 0.221, reflecting overall probabilistic accuracy of the predictions in this cross-sectional setting. Application in future survey rounds or specific provincial contexts will require recalibration because baseline readiness and risk distributions are likely to differ [42]. Decision-use is more appropriately evaluated with net-benefit methods such as decision-curve analysis, which compare alternative strategies (for example, universal support, no targeting, or prioritisation of high-risk deciles) across thresholds aligned to programme capacity [52].

Strengths and limitations of this study

A strength of this study is the analytical framework used to evaluate predictive performance within a nationally sampled cohort. Clustered cross-validation at the early learning programme level reduced the risk of information leakage arising from shared environments, and predictors used to derive the readiness outcome were excluded to avoid circularity between covariates and the classification target.

Domain and source ablation analyses further enabled evaluation of the relative contribution of child, caregiver, and programme variables to predictive performance. These features strengthen interpretation of the findings and align with current reporting standards for prediction studies, including transparent specification of model development, validation strategy, and performance metrics [42,43].

This study is not without limitations. The cross-sectional design does not permit temporal ordering and therefore precludes causal inference. The use of categorical readiness bands also reduces information relative to continuous developmental scores and may attenuate discriminatory ability near classification thresholds [42]. Caregiver data were unavailable for approximately one quarter of children, which may have reduced the explanatory power of household variables and could introduce selection bias if missingness was systematic, although the direction and magnitude of this bias cannot be determined without additional analysis.

In addition, although the LPQA instrument includes domains intended to capture classroom process quality—such as teaching strategies, planning and assessment, and relationships and interactions—programme-level predictive signal may be attenuated when models simultaneously include child-, household-, and socioeconomic factors that account for a large proportion of variation in developmental outcomes, a pattern widely documented in early childhood research [6]. Finally, model performance was evaluated using internal cross-validation only. Given the clustered sampling design and potential regional heterogeneity in early learning environments, external validation in independent samples or across provinces would be necessary before applying these models to operational targeting frameworks.

Overall, within an adversity-defined cohort at school entry, predictive variation in early learning readiness was structured primarily by maturational timing, growth-related indicators, and caregiver-linked resources, with limited incremental signal from programme descriptors within the measured feature set. This pattern is consistent with developmental systems models in which early learning reflects the interaction of biological status and caregiving environments [21,30], and highlights a measurement gap in indicators of enacted pedagogical quality. Stratification approaches based on routinely collected child and caregiver information may therefore help identify children requiring additional assessment or support, provided thresholds are aligned to programme capacity and model performance is evaluated across socioeconomic strata to avoid widening inequities [42].

Acknowledgements

I would like to thank Mr. Devon Jarvis, from the MIND Institute at the University of the Witwatersrand, for advising on the machine learning architecture.

Ethics and Consent

JEthical approval was granted by the University of Cape Town (HREC 23/021) and the University of the Witwatersrand (M230551). Procedures adhered to South Africa's Protection of Personal Information Act; public use data are released under a restricted licence that confines analysis to the approved protocol [24]. Written informed consent from caregivers and age appropriate assent from children were obtained before assessment.

Funding Information

M.L. is supported by a postdoctoral fellowship from the National Research Fund, South Africa.

ORCIDs

Michelle Leal: 0000-0001-8152-6765

References

Black MM, Walker SP, Fernald LCH, Andersen CT, DiGirolamo AM, Lu C, et al. Early childhood development coming of age: science through the life course. *The Lancet 2017 Jan389*(10064):77–90. **doi:10.1016/S0140-6736(16)31389-7**.

McCoy DC, Peet ED, Ezzati M, Danaei G, Black MM, Sudfeld CR, et al. Early Childhood Developmental Status in Low- and Middle-Income Countries: National, Regional, and Global Prevalence Estimates Using Predictive Modeling. *PLOS Medicine 2016 June 713*(6):e1002034. **doi:10.1371/journal.pmed.1002034.**

Lu C, Cuartas J, Fink G, McCoy D, Liu K, Li Z, et al. Inequalities in early childhood care and development in low/middle-income countries: 2010–2018. *BMJ Global Health 2020 Feb5*(2):e002314. **doi:10.1136/bmjgh-2020-002314.**

Sudfeld CR, Charles McCoy D, Danaei G, Fink G, Ezzati M, Andrews KG, et al. Linear Growth and Child Development in Low- and Middle-Income Countries: A Meta-Analysis. *Pediatrics 2015 May 1135*(5):e1266–75. **doi:10.1542/peds.2014-3111**.

Mashburn AJ, Pianta RC, Hamre BK, Downer JT, Barbarin OA, Bryant D, et al. Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills. *Child Development 2008 May 179*(3):732–49. **doi:10.1111/j.1467-8624.2008.01154.x.**

Ulferts H, Wolf KM, Anders Y. Impact of Process Quality in Early Childhood Education and Care on Academic Outcomes: Longitudinal Meta-Analysis. *Child Development 2019 Sept 190(*5):1474–89. **doi:10.1111/cdev.13296**.

Wolf S, Aber JL, Behrman JR, Peele M. Longitudinal causal impacts of preschool teacher training on Ghanaian children's school readiness: Evidence for persistence and fade-out. *Developmental Science 2019 Sept22*(5):e12878. **doi:10.1111/desc.12878**.

Baker-Henningham H, Bowers M, Francis T, Vera-Hernández M, Walker SP. The Irie Classroom Toolbox, a universal violence-prevention teacher-training programme, in Jamaican preschools: a single-blind, cluster-randomised controlled trial. *The Lancet Global Health 2021 Apr9*(4):e456–68. **doi:10.1016/S2214-109X(21)00002-4**.

World Bank. *The State of Global Learning Poverty: 2022 Update*. 2022.

Richter LM, Daelmans B, Lombardi J, Heymann J, Boo FL, Behrman JR, et al. Investing in the foundation of sustainable development: pathways to scale up for early childhood development. *The Lancet 2017 Jan389*(10064):103–18. **doi:10.1016/S0140-6736(16)31698-1.**

UNICEF; World Health Organization; World Bank Group. *Levels and trends in child malnutrition: Key findings of the 2023 edition* (Joint Child Malnutrition Estimates). 2023.

UNESCO. UNESCO Global Education Monitoring Report. Monitoring SDG 4: Early childhood care and education. 2024. Available from: **https://www.unesco.org/gem-report/en/ecce**.

Mullis IVS, Martin MO, Foy P, Stanco GM. *Progress in International Reading Literacy Study* (PIRLS) 2021 International Results in Reading. 2023.

World Bank. The Human Capital Index 2020 Update: Human Capital in the Time of COVID-19. 2020.

World Bank. Learning-Adjusted Years of Schooling (LAYS), indicator HD.HCI.LAYS. 2024.

DataDrive2030; Department of Basic Education; Thrive by Five Partners. *Thrive by Five Index report* (revised Aug 2022). 2022.

Britto PR, Lye SJ, Proulx K, Yousafzai AK, Matthews SG, Vaivada T, et al. Nurturing care: promoting early childhood development. *The Lancet 2017 Jan389*(10064):91–102. **doi:10.1016/S0140-6736(16)31390-3.**

World Health Organization, United Nations Children's Fund. *Nurturing care for early childhood development: a framework for helping children survive and thrive to transform health and human potential*. 2018.

Masten AS. Ordinary magic: Resilience processes in development. *American Psychologist 200156*(3):227–38. **doi:10.1037//0003-066x.56.3.227**.

Masten AS, Lucke CM, Nelson KM, Stallworthy IC. Resilience in Development and Psychopathology: Multisystem Perspectives. *Annual Review of Clinical Psychology 2021 May 717*(1):521–49. **doi:10.1146/annurev-clinpsy-081219-120307**.

Black MM, Walker SP, Fernald LCH, Andersen CT, DiGirolamo AM, Lu C, et al. Early childhood development coming of age: science through the life course. *The Lancet 2017 Jan389*(10064):77–90. **doi:10.1016/S0140-6736(16)31389-7**.

Requejo J, Strong K, Agweyu A, Billah SM, Boschi-Pinto C, Horiuchi S, et al. Measuring and monitoring child health and wellbeing: recommendations for tracking progress with a core set of indicators in the Sustainable Development Goals era. *The Lancet Child & Adolescent Health 2022 May 6*(5):345–52. **doi:10.1016/S2352-4642(22)00039-6**.

DataDrive2030 Technical Team. *South Africa 2024* Thrive by Five Index – Sampling strategy. 2024.

DataDrive2030 Technical Team. South Africa 2024 Thrive by Five Index – datasets summary: enrolled children. 2025.

DataDrive2030. ELOM Social-Emotional Rating Scale: technical manual and rating sheet. 2022. Available from: **https://datadrive2030.co.za/wp-content/uploads/2022/11/ELOM-Social-Emotional-Rating-Scale_3.pdf.**

Biersteker L, Tredoux C, Mattes F, Dawes A. Report on the development of the ELOM Learning Programme Quality Assessment Tool. 2022.

World Health Organization Multicentre Growth Reference Study Group. WHO Child Growth Standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development. 2006.

Anderson KJ, Henning TJ, Moonsamy JR, Scott M, Du Plooy C, Dawes ARL. Test–retest reliability and concurrent validity of the South African Early Learning Outcomes Measure (ELOM). *South African Journal of Childhood Education 2021 June 1711*(1). Available from: **http://www.sajce.co.za/index.php/sajce/article/view/881. doi:10.4102/sajce.v11i1.881**.

Perkins JM, Kim R, Krishna A, McGovern M, Aguayo VM, Subramanian SV. Understanding the association between stunting and child development in low- and middle-income countries: Next steps for research and intervention. *Social Science & Medicine 2017 Nov193*:101–9. **doi:10.1016/j.socscimed.2017.09.039**.

Grantham-McGregor S, Cheung YB, Cueto S, Glewwe P, Richter L, Strupp B. Developmental potential in the first 5 years for children in developing countries. *The Lancet 2007 Jan369*(9555):60–70. **doi:10.1016/S0140-6736(07)60032-4**.

Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ 2024 Apr 16*:e078378. **doi:10.1136/bmj-2023-078378**.

Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ 2024 Jan* 15384:e074820. **doi:10.1136/bmj-2023-074820**.

Shmueli G. To Explain or to Predict? *Statistical Science 2010 Aug 125*(3). **Available from: https://projecteuclid.org/journals/statistical-science/volume-25/issue-3/To-Explain-or-to-Predict/10.1214/10-STS330.full. doi:10.1214/10-STS330**.

MacNell N, Feinstein L, Wilkerson J, Salo PM, Molsberry SA, Fessler MB, et al. Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting. *PLOS ONE 2023 Jan 1318*(1):e0280387. **doi:10.1371/journal.pone.0280387**.

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive Mixtures of Local Experts. *Neural Computation 1991 Feb3*(1):79–87. **doi:10.1162/neco.1991.3.1.79**.

Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *Proceedings of the International Conference on Learning Representations 2017*. Available from: **https://doi.org/10.48550/arXiv.1701.06538. doi:10.48550/arXiv.1701.06538**.

Masoudnia S, Ebrahimpour R. Mixture of experts: a literature survey. *Artificial Intelligence Review 2014 Aug42*(2):275–93. **doi:10.1007/s10462-012-9338-y**.

Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology 2019 June*110:12–22. **doi:10.1016/j.jclinepi.2019.02.004**.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems (NeurIPS) 201730:3146–54.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. B*ioinformatics 2001 June 117*(6):520–5. **doi:10.1093/bioinformatics/17.6.520**.

Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. 2017. Available from: **https://arxiv.org/abs/1705.07874. doi:10.48550/ARXIV.1705.07874**.

Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2019. **Available from: http://link.springer.com/10.1007/978-3-030-16399-0. doi:10.1007/978-3-030-16399-0**.

Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Annals of Internal Medicine 2015 Jan 6162*(1):55–63. **doi:10.7326/M14-0697**.

Urruticoechea A, Oliveri A, Vernazza E, Giménez-Dasí M, Martínez-Arias R, Martín-Babarro J. The Relative Age Effects in Educational Development: A Systematic Review. *International Journal of Environmental Research and Public Health 2021 Aug 2618*(17):8966. **doi:10.3390/ijerph18178966**.

Lu C, Cuartas J, Fink G, McCoy D, Liu K, Li Z, et al. Inequalities in early childhood care and development in low/middle-income countries: 2010–2018. *BMJ Global Health 2020 Feb5*(2):e002314. **doi:10.1136/bmjgh-2020-002314**.

McCoy DC, Waldman M, Fink G. Measuring early childhood development at a global scale: Evidence from the Caregiver-Reported Early Development Instruments. *Early Childhood Research Quarterly 2018* 3445:58–68. **doi:10.1016/j.ecresq.2018.05.002**.

Weber A, Darmstadt GL, Rao N. Gender disparities in child development in the east Asia-Pacific region: a cross-sectional, population-based, multicountry observational study. *The Lancet Child & Adolescent Health 2017 Nov1*(3):213–24. **doi:10.1016/S2352-4642(17)30073-1**.

Vermeer HJ, Van IJzendoorn MH, Cárcamo RA, Harrison LJ. Quality of Child Care Using the Environment Rating Scales: A Meta-Analysis of International Studies. *International Journal of Early Childhood 2016 Apr48*(1):33–60. **doi:10.1007/s13158-015-0154-9**.

Gladstone MJ, Chandna J, Kandawasvika G, Ntozini R, Majo FD, Tavengwa NV, et al. Independent and combined effects of improved water, sanitation, and hygiene (WASH) and improved complementary feeding on early neurodevelopment among children born to HIV-negative mothers in rural Zimbabwe: Substudy of a cluster-randomized trial. *PLOS Medicine 2019 Mar 2116*(3):e1002766. **doi:10.1371/journal.pmed.1002766**.

Burchinal M. Measuring Early Care and Education Quality. *Child Development Perspectives 2018 Mar 112*(1):3–9. **doi:10.1111/cdep.12260**.

Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making 2006 Nov26*(6):565–74. **doi:10.1177/0272989X06295361**

Tables

Table 1. Sample characteristics (Total N = 2,278)

Panel	Level	Total	On Track	Not On Track	P-value
Panel A. Child Demographic & Biological Factors					
Child age (months)	0.16	54.8 (2.4), n=2278	55.5 (2.3), n=872	54.4 (2.4), n=1406	<0.001
Sex	Female	1184 (50.9%)	507 (55.9%)	677 (48.0%)	0.009
Height-for-Age (HAZ Score)		-1.1 (0.9), n=2271	-1.1 (0.9), n=873	-1.1 (0.9), n=1398	0.869
Stunting Status (HAZ < -2)	Stunted	333 (14.2%)	87 (10.9%)	246 (16.2%)	0.014
Number of languages spoken at home	More than one	72 (2.9%)	35 (4.1%)	37 (2.1%)	0.043
Assessment matches home language	Yes	2277 (99.9%)	876 (100.0%)	1401 (99.9%)	0.285
Panel B. Household Characteristics & Home Environment					
Household Assets Score (0-10)		5.2 (1.9), n=1940	5.4 (2.0), n=764	5.1 (1.9), n=1176	0.035
Caregiver Education	Primary or less	171 (9.1%)	54 (8.3%)	117 (9.6%)	0.643
	Secondary	1177 (61.9%)	442 (61.4%)	735 (62.2%)	
	Tertiary	592 (29.0%)	268 (30.3%)	324 (28.2%)	
Caregiver Employment	Employed	896 (41.3%)	397 (45.7%)	499 (38.5%)	0.031

Panel	Level	Total	On Track	Not On Track	P-value
Children's books in home	1-5	1206 (62.1%)	499 (66.9%)	707 (59.1%)	0.002
	6+	154 (6.5%)	83 (8.3%)	71 (5.5%)	
	None	572 (31.3%)	176 (24.8%)	396 (35.5%)	
Has Road to Health Booklet	Yes	1871 (97.0%)	747 (98.5%)	1124 (96.1%)	0.032
Panel C1. Programme Context & Administration					
Province	Eastern Cape	286 (13.4%)	91 (12.2%)	195 (14.1%)	0.282
	Free State	248 (6.8%)	124 (9.9%)	124 (4.9%)	
	Gauteng	419 (18.1%)	141 (16.3%)	278 (19.1%)	
	KwaZulu-Natal	316 (17.3%)	107 (15.5%)	209 (18.3%)	
	Limpopo	301 (23.2%)	134 (23.9%)	167 (22.8%)	
	Mpumalanga	173 (8.0%)	73 (8.4%)	100 (7.7%)	
	North West	175 (5.6%)	74 (6.2%)	101 (5.2%)	
	Northern Cape	183 (1.7%)	70 (1.8%)	113 (1.6%)	
	Western Cape	179 (6.1%)	62 (5.8%)	117 (6.3%)	

Panel	Level	Total	On Track	Not On Track	P-value
Programme Location	Community/Municipal	176 (7.5%)	60 (7.4%)	116 (7.5%)	0.928
	ECD centre	1375 (73.1%)	536 (73.7%)	839 (72.7%)	
	Home-based	558 (19.5%)	197 (18.8%)	361 (19.8%)	
Programme Registration Status	Registered	1548 (72.6%)	634 (75.6%)	914 (70.8%)	0.210
Fee level (Low/Medium/High)	High	224 (5.0%)	136 (7.9%)	88 (3.2%)	0.004
	Low	1158 (58.6%)	404 (53.9%)	754 (61.4%)	
	Medium	898 (36.4%)	336 (38.2%)	562 (35.4%)	
Principal Education (3 levels)	Primary or less	72 (3.0%)	22 (2.4%)	50 (3.4%)	0.503
	Secondary	671 (27.0%)	237 (25.8%)	434 (27.7%)	
	Tertiary	1537 (70.0%)	617 (71.7%)	920 (69.0%)	
Years as Principal		12.1 (8.4), n=2274	11.5 (8.0), n=875	12.4 (8.6), n=1399	0.197

Panel	Level	Total	On Track	Not On Track	P-value
Panel C2. Facility Structural Quality & WASH					
Facility has Mains Electricity	Yes	1760 (77.3%)	703 (77.8%)	1057 (77.1%)	0.796
Secured perimeter (Fence & Gate)	Yes	2141 (96.0%)	838 (96.4%)	1303 (95.8%)	0.655
Outdoor play area	Nearby	8 (0.3%)	4 (0.4%)	4 (0.2%)	0.367
	No	327 (11.2%)	95 (10.0%)	232 (12.0%)	
	On premises	1945 (88.5%)	777 (89.6%)	1168 (87.8%)	
Panel C3. Institutional Process Quality (LPQA)					
Total LPQA Score		34.3 (13.0), n=2278	34.4 (12.6), n=872	34.2 (13.1), n=1406	0.825
Staff-Child Interaction Score		7.8 (2.3), n=2278	7.8 (2.3), n=872	7.7 (2.3), n=1406	0.859
Class Size (Children Present)		20.0 [15.0, 28.0], n=2269	21.9 (10.8), n=874	20.0 [15.0, 29.0], n=1395	0.790
Panel D. Educator Human Capital & Pedagogical Beliefs					
Practitioner trained in last 12m	Yes	1035 (52.3%)	403 (52.8%)	632 (52.0%)	0.833
Practitioner salary (Collapsed)	< R1500	323 (39.1%)	132 (39.2%)	191 (39.1%)	0.346
	R1500–R2999	334 (39.4%)	130 (36.7%)	204 (41.3%)	
	R3000+	235 (21.5%)	118 (24.1%)	117 (19.6%)	
Curriculum Awareness (NCF)	Yes	788 (80.4%)	347 (81.6%)	441 (79.5%)	0.569
Practitioner Mindset (Growth vs Fixed)	Growth	461 (91.8%)	198 (92.1%)	263 (91.6%)	0.844

Notes: Estimates are design-based and weighted. Adversity defined by growth risk (HAZ < -1) or lowest asset quintile. Target-derived variables (ELOM/SEF scores) were excluded from the covariate set. P-values compare children on track for early learning versus children not on track for early learning.

Table 2. Model performance for prediction of early learning status in the adversity sub-sample

Model	AUC, mean (SD)	AP, mean (SD)	Log loss, mean (SD)
Histogram gradient boosting	0.66 (0.02)	0.55 (0.01)	0.64 (0.01)
Mixture-of-experts	0.62 (0.02)	0.51 (0.02)	0.65 (0.01)
Elastic-net logistic regression	0.60 (0.03)	0.49 (0.04)	1.07 (0.10)

Notes: Values are mean (SD) across five-fold stratified group cross-validation with clustering at the early learning programme level. Analyses were restricted to children meeting the predefined adversity criteria. The Mixture-of-Experts neural network utilised a three-subnetwork architecture. Histogram-based gradient boosting and elastic-net logistic regression were fitted as benchmark models using the same cross-validation folds. Models were trained and evaluated without sampling weights. AUC = area under the receiver operating characteristic curve; AP = average precision.

Figures

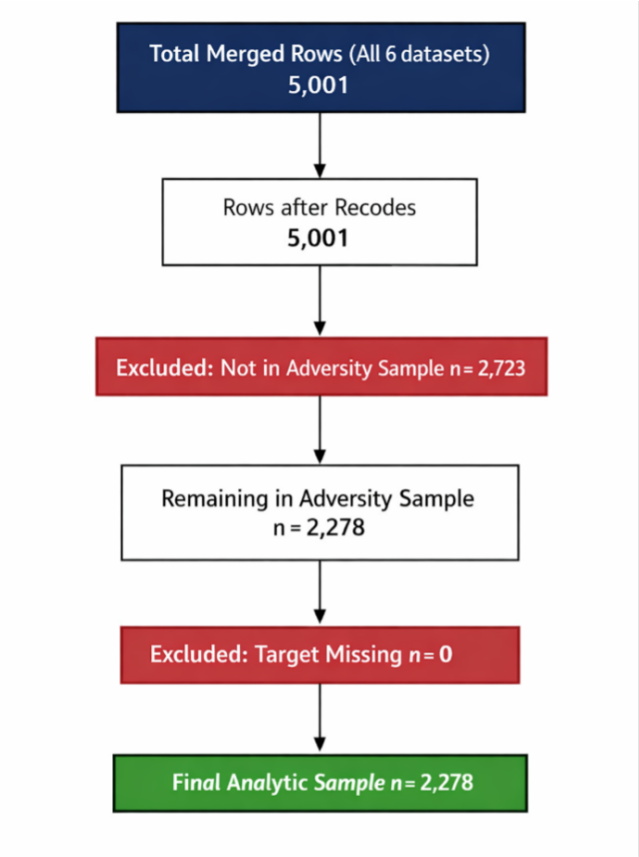
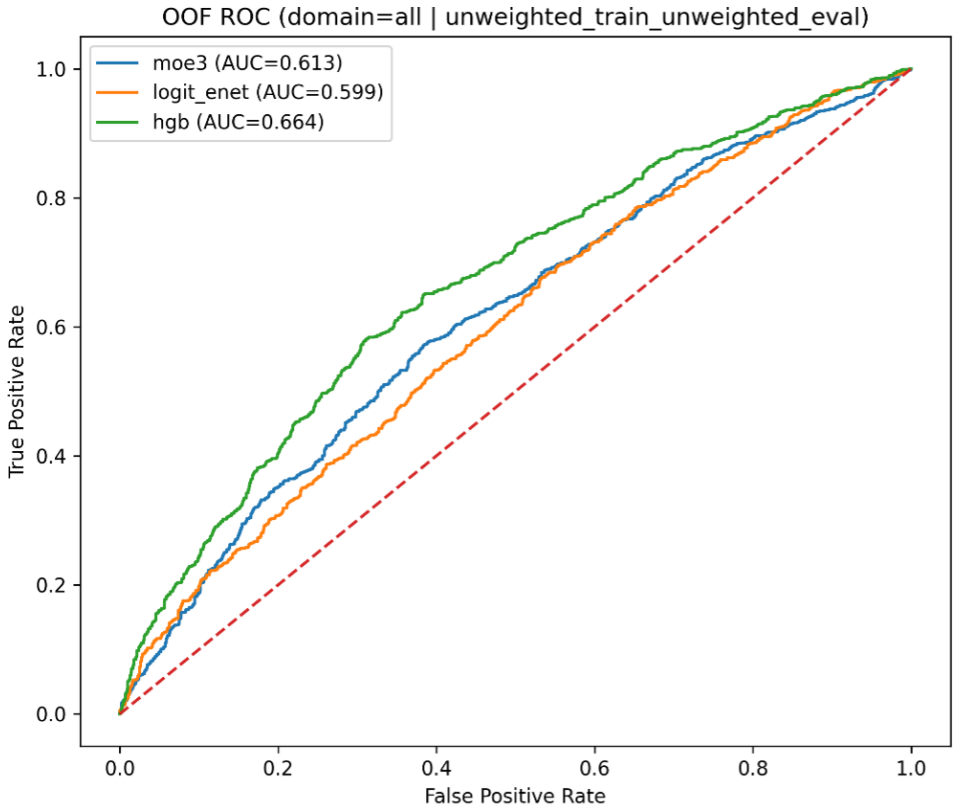


Figure 1. Study sample flow diagram.

Of 5,001 merged records across six datasets, 2,723 were excluded because they were not in the adversity sample. No observations were excluded due to missing outcome data. The final analytic sample comprised 2,278 children.



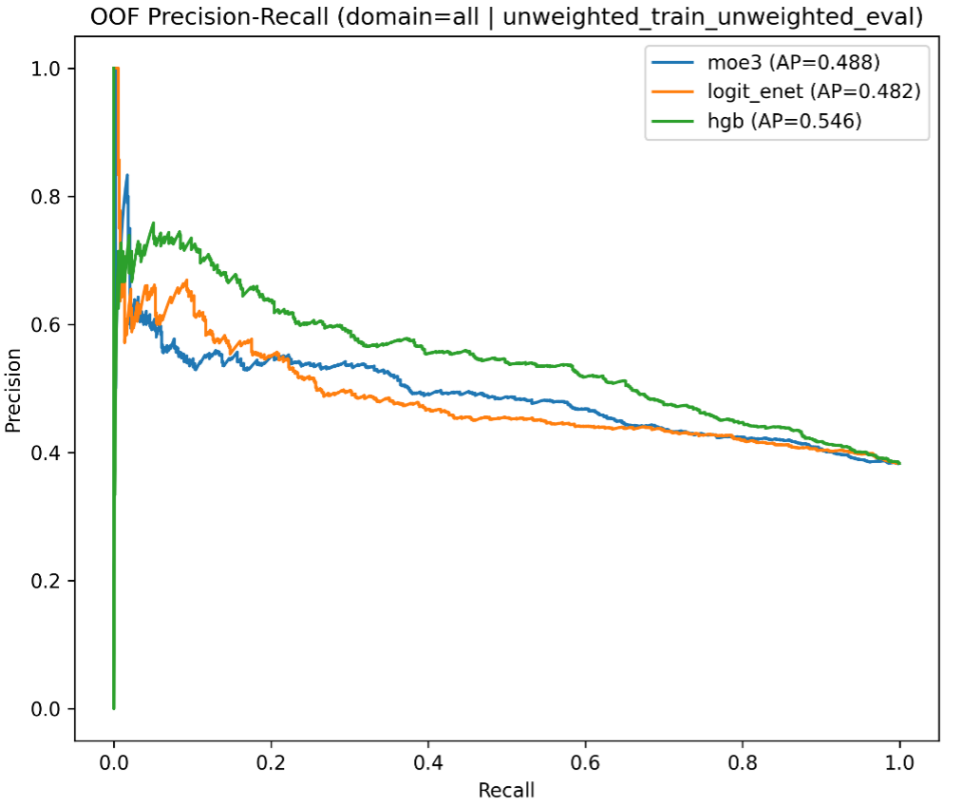


Figure 2

Cross-validated discrimination for prediction of early learning status in the adversity sub-sample.

(A) Receiver operating characteristic curves and (B) precision–recall curves for the histogram-based gradient boosting (HGB), Mixture-of-Experts (MoE), and elastic-net logistic regression models. Curves are based on predictions obtained from five-fold stratified group cross-validation with clustering at the early learning programme level. AUC = area under the receiver operating characteristic curve; AP = average precision.

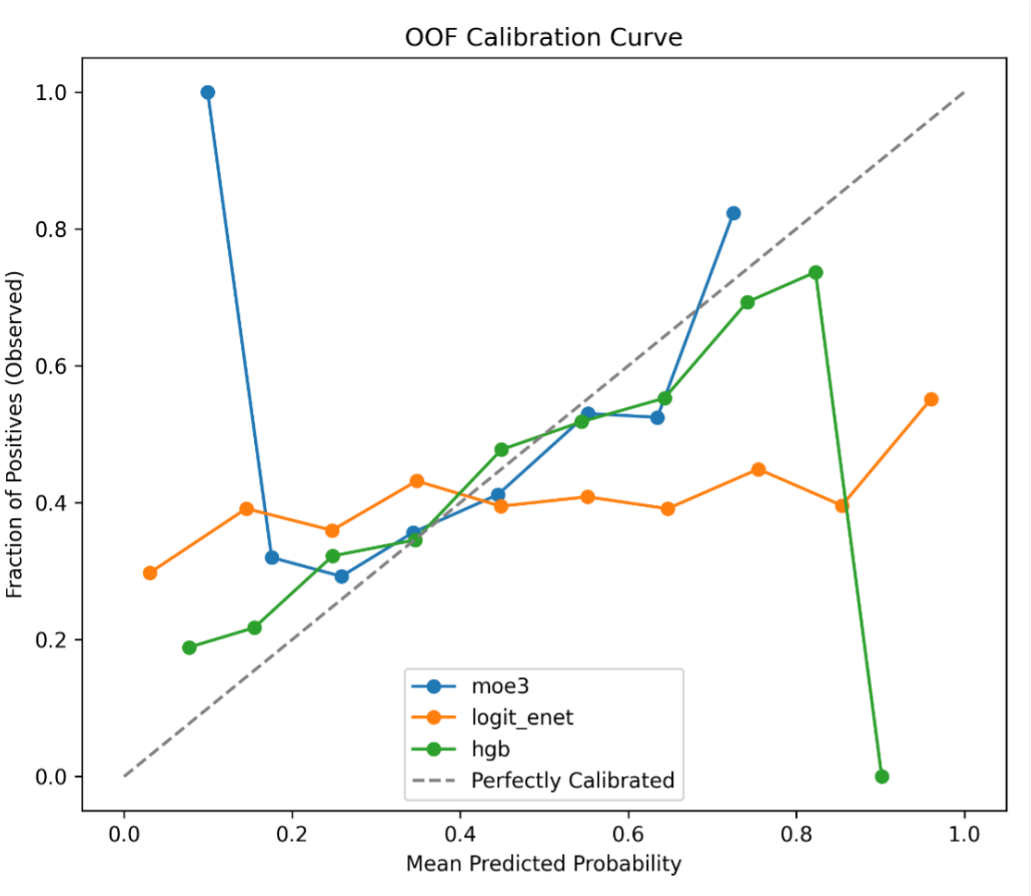


Figure 3

Cross-validated calibration curves for prediction of early learning status in the adversity sub-sample.

Calibration curves compare mean predicted probabilities with observed outcome frequencies across deciles of predicted risk for the histogram-based gradient boosting (HGB), Mixture-of-Experts (MoE), and elastic-net logistic regression models. The dashed line represents perfect calibration. Curves are based on predictions obtained from five-fold stratified group cross-validation with clustering at the early learning programme level.

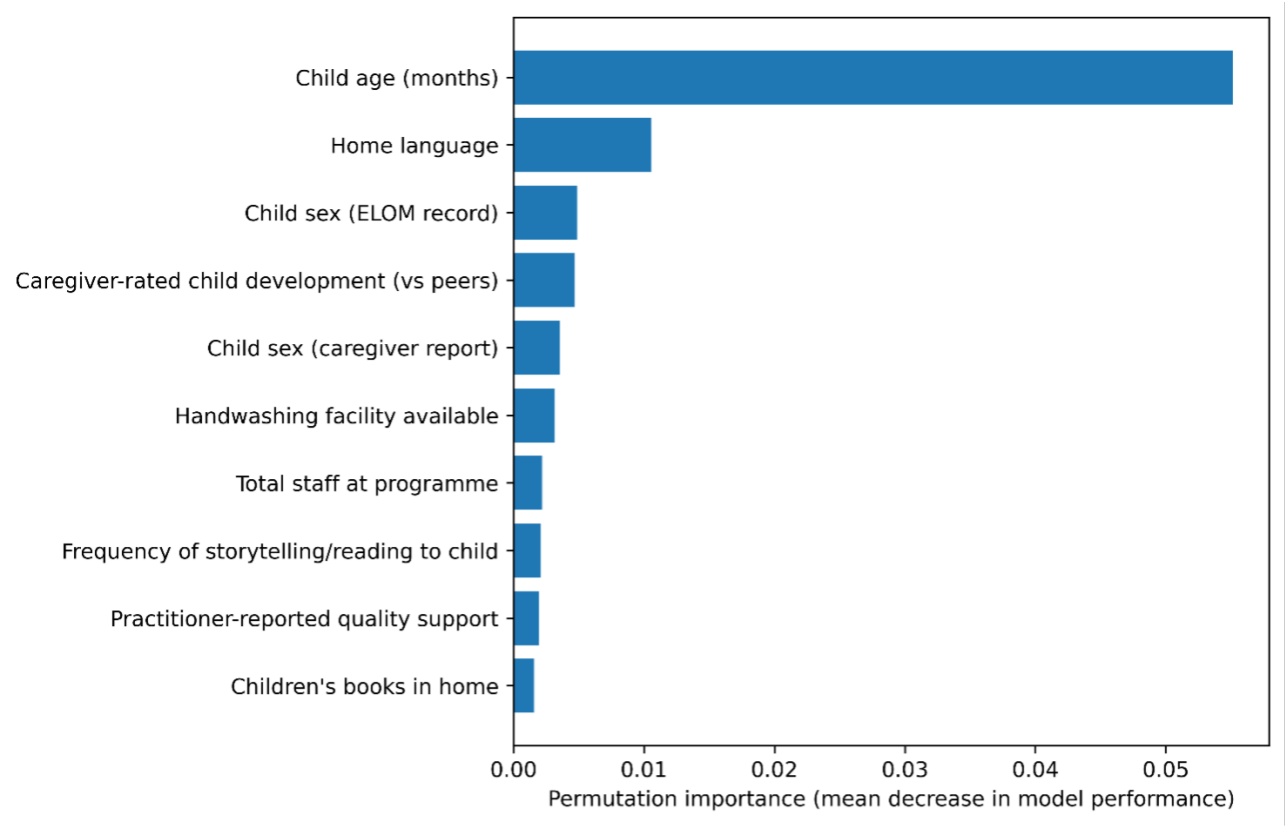


Figure 4

Permutation importance of predictors in the histogram-based gradient boosting model.

Permutation importance values represent the mean decrease in model performance when each predictor is randomly permuted, indicating the relative contribution of predictors to model discrimination in the adversity sub-sample. Child age was the dominant predictor, followed by home language and child sex indicators. Remaining predictors represent contextual caregiver, programme, and household characteristics.

Appendices

Supplementary Tables

Table S1. Domain ablation analyses for prediction of early learning on-track status in the adversity sub-sample

Domain	Analysis	AUC, mean (SD)	ΔAUC vs full	AP, mean (SD)	Log loss, mean (SD)
Caregiver	DROP	0.62 ± 0.03	0.00	0.50 ± 0.03	0.66 ± 0.03
Child	DROP	0.61 ± 0.02	−0.02	0.49 ± 0.03	0.67 ± 0.03
ELP	DROP	0.64 ± 0.02	+0.01	0.52 ± 0.02	0.65 ± 0.01
Other	DROP	0.62 ± 0.02	0.00	0.51 ± 0.02	0.65 ± 0.01
Caregiver	ONLY	0.59 ± 0.03	−0.03	0.48 ± 0.03	0.66 ± 0.01
Child	ONLY	0.63 ± 0.01	+0.01	0.52 ± 0.02	0.64 ± 0.00
ELP	ONLY	0.59 ± 0.02	−0.03	0.48 ± 0.02	0.67 ± 0.02

Notes: Values are mean ± SD from five-fold stratified group cross-validation with clustering at the early learning programme level. Analyses were restricted to the adversity sub-sample. "DROP" indicates model performance after removal of predictors from the specified domain. "ONLY" indicates models trained using predictors from the specified domain alone. ΔAUC represents the change in discrimination relative to the full Mixture-of-Experts model (AUC = 0.62). Models were trained and evaluated without sampling weights. AUC = area under the receiver operating characteristic curve; AP = average precision.

Supplementary Tables

Table S2. Source-restricted analyses for prediction of early learning on-track status in the adversity sub-sample

Source	Analysis	AUC (mean ± SD)	ΔAUC vs full	AP (mean ± SD)	Log loss (mean ± SD)
Caregiver	DROP	0.62 ± 0.03	0.00	0.50 ± 0.03	0.66 ± 0.03
Child	DROP	0.62 ± 0.02	0.00	0.51 ± 0.02	0.65 ± 0.01
Facility	DROP	0.62 ± 0.01	0.00	0.50 ± 0.02	0.65 ± 0.01
LPQA	DROP	0.61 ± 0.03	−0.01	0.49 ± 0.03	0.66 ± 0.01
Practitioner	DROP	0.63 ± 0.02	+0.01	0.52 ± 0.02	0.65 ± 0.01
Principal	DROP	0.62 ± 0.02	0.00	0.51 ± 0.01	0.66 ± 0.02
Caregiver	ONLY	0.59 ± 0.03	−0.03	0.48 ± 0.03	0.66 ± 0.01
Facility	ONLY	0.54 ± 0.02	−0.08	0.41 ± 0.02	0.68 ± 0.03
LPQA	ONLY	0.59 ± 0.01	−0.03	0.48 ± 0.03	0.66 ± 0.00
Practitioner	ONLY	0.52 ± 0.03	−0.10	0.41 ± 0.02	0.67 ± 0.01
Principal	ONLY	0.59 ± 0.02	−0.04	0.47 ± 0.02	0.66 ± 0.01

Notes: Values are mean ± SD from five-fold stratified group cross-validation with clustering at the early learning programme level. Analyses were restricted to the adversity sub-sample. "DROP" indicates model performance after removal of predictors from the specified source within the full feature set. "ONLY" indicates models trained using predictors from the specified source alone. ΔAUC represents the change in discrimination relative to the full Mixture-of-Experts model (AUC = 0.62). Machine-learning models were trained and evaluated without sampling weights. AUC = area under the receiver operating characteristic curve; AP = average precision.

Supplementary Table S3. Sensitivity of adversity subgroup composition to alternative HAZ thresholds (-1.0, -1.5, and -2.0 SD)

HAZ threshold (SD)	HAZ below threshold, n (%)	Lowest weighted asset quintile, n (%)	Adversity subgroup(low asset quintile and/or HAZ <threshold), n (%)	Overlap (low asset quintile and HAZ <threshold), n (%)	On track in adversity, n/N (%)
-1.0	1587 (31.7)	1128 (22.6)	2278 (45.6)	437 (8.7)	872/2278 (38.3)
-1.5	796 (15.9)	1128 (22.6)	1702 (34.0)	222 (4.4)	619/1702 (36.4)
-2.0	331 (6.6)	1128 (22.6)	1362 (27.2)	97 (1.9)	475/1362 (34.9)

Note. HAZ = height-for-age z-score (ZHFA). The lowest asset quintile was operationalised as the weighted bottom 20% of the household asset count distribution using caregiver sampling weights (weight_pcg). Adversity in this sensitivity analysis was defined as membership in the lowest weighted asset quintile and/or HAZ below the specified threshold. On-track early learning status was defined as on-track ELOM performance among children classified in the adversity subgroup with observed outcome data. Percentages are calculated using the full recoded sample denominator (N=5001), except on track percentages, which use the adversity subgroup denominator at each threshold. The HAZ < -1.0 row reproduces the primary adversity definition used in the main analysis.

Supplementary Tables

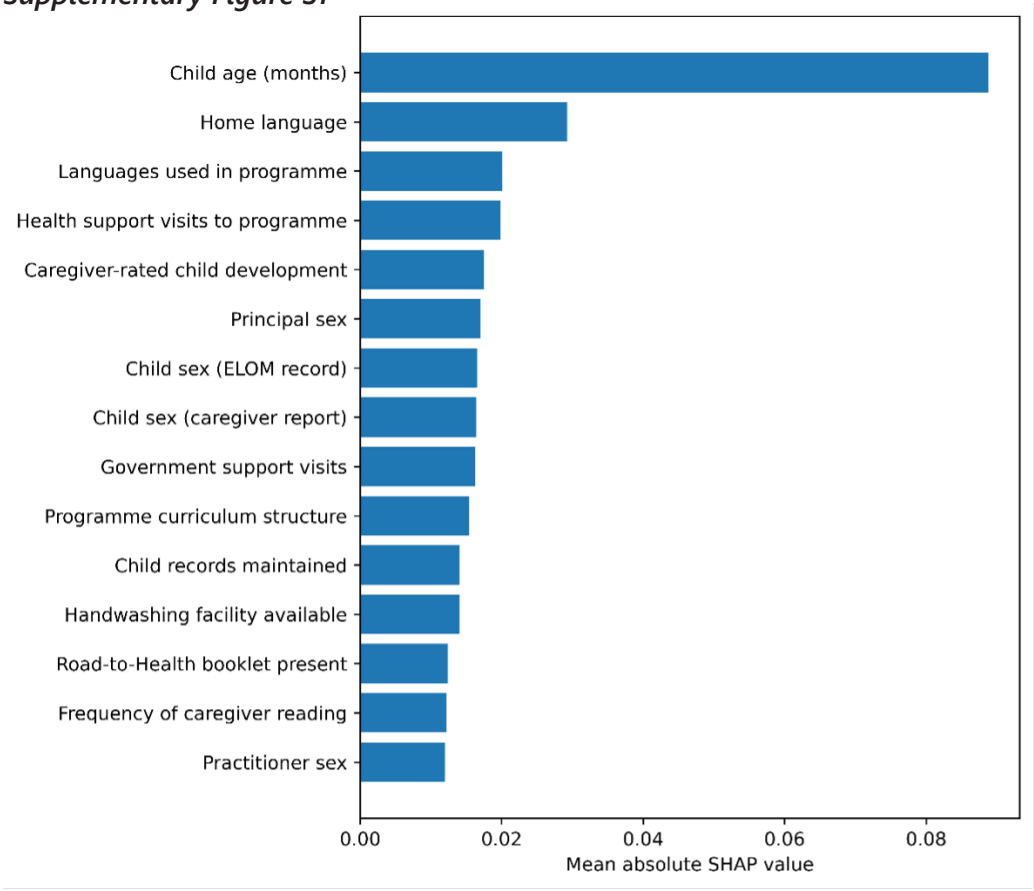
Supplementary Table S4. Sensitivity analysis of sampling-weight specifications on model performance

Weighting specification	Model	AUC (mean ± SD)	Average precision (mean ± SD)	Log loss (mean ± SD)
Unweighted training, unweighted evaluation	Histogram-based gradient boosting	0.664 ± 0.015	0.552 ± 0.013	0.636 ± 0.013
	Mixture-of-Experts	0.623 ± 0.017	0.508 ± 0.016	0.652 ± 0.009
	Elastic-net logistic regression	0.600 ± 0.031	0.488 ± 0.035	1.071 ± 0.096
Unweighted training, weighted evaluation	Histogram-based gradient boosting	0.639 ± 0.051	0.505 ± 0.058	0.650 ± 0.031
	Mixture-of-Experts	0.590 ± 0.040	0.466 ± 0.067	0.663 ± 0.017
	Elastic-net logistic regression	0.591 ± 0.046	0.469 ± 0.055	1.075 ± 0.095
Weighted training, weighted evaluation	Histogram-based gradient boosting	0.600 ± 0.025	0.487 ± 0.040	0.678 ± 0.016
	Mixture-of-Experts	0.583 ± 0.043	0.449 ± 0.038	0.658 ± 0.007
	Elastic-net logistic regression	0.587 ± 0.037	0.467 ± 0.058	1.118 ± 0.075

Note. Model performance was estimated using five-fold stratified group cross-validation with grouping at the ECD-centre level. Three weighting specifications were evaluated: (i) unweighted training and evaluation (primary analysis), (ii) unweighted training with weighted evaluation metrics, and (iii) weighted training and weighted evaluation. Performance metrics represent means and standard deviations across folds.

Supplementary Figures

Supplementary Figure S1



SHAP feature importance for the histogram-based gradient boosting model. Mean absolute SHAP (SHapley Additive exPlanations) values for the top 15 predictors contributing to model predictions of early learning status in the adversity sub-sample. Higher values indicate greater influence on model predictions.

DATADRIVE 2030

Learn more at **DataDrive2030.co.za** | Follow us on LinkedIn **@DataDrive2030**