



JUNE 2026



Florida's Early Childhood
Integrated Data System

Technical Brief: Children's Academic Outcomes Associated with Early Child Care and Education System Participation

FISCAL YEAR 2025-2026

EARLY CHILDHOOD POLICY RESEARCH GROUP (ECPRG)

Phillip Sherlock, Ph.D., Herman T. Knopf, Ph.D., Jiawei Li, M.S., Robert Chapman

The Sunshine State Early Childhood Information Portal Project described was supported by the Preschool Development Grant Birth through Five Initiative (PDG B-5) Grant Number 90TP0068-03-02 from the Office of Child Care, Administration for Children and Families, U.S. Department of Health and Human Services. Authority and Funding is Section 46 of Chapter 2023-239, Laws of Florida, from the American Rescue Plan (ARP) Act Discretionary Child Care and Development Block Grant Trust Fund. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Office of Child Care, the Administration for Children and Families, or the U.S. Department of Health and Human Services.





Early Childhood Policy Research Group (ECPRG)

This work is the result of the ECPRG; a collaborative team housed at the University of Florida Anita Zucker Center for Excellence in Early Childhood Studies that consists of the following partnership organizations and individuals.

Anita Zucker Center for Excellence in Early Childhood Studies

University of Florida

- Herman T. Knopf, Ph.D.
- Phillip Sherlock, Ph.D.
- Dévonja Daley
- Jiawei Li, M.S.
- Rachel Eubanks, M.Ed.
- Mary Kay Rodgers, Ph.D.
- Robert Chapman
- Daveyon Brown
- Xiaomeng Xiong, M.A.

Department of Medical Social Sciences

Northwestern University, Feinberg School of Medicine

- Maxwell Mansolf, Ph.D.

Department of Health Outcomes & Biomedical Informatics

University of Florida

- Elizabeth Shenkman, Ph.D.
- Deepa Ranka, M.S.
- Erik Schmidt

Table of Contents

Children’s Academic Outcomes Associated with Early Child Care and Education System Participation	5
About This Research.....	5
The 12 Groups of VPK Children	6
What Two Years of Research Tell Us	7
A Strengthened Body of Evidence.....	9
Implications for the DEL	10
Closing Thoughts.....	12
Technical Appendix	13

Children's Academic Outcomes Associated with Early Child Care and Education System Participation

This report brings together two years of research on children enrolled in Florida's Voluntary Prekindergarten (VPK) program. The findings from our current cluster analysis (FY 2025-26) build directly on, and are strongly supported by, the ecological study published last year (FY 2024-25) by the ECPRG at the University of Florida.

Together, the two studies tell a consistent and compelling story: VPK works, especially for the children who need it most. But whether children realize those benefits depends heavily on who they are, what they bring with them, and whether they show up consistently.

TWO YEARS, ONE STORY

The 2024-25 study showed that children who start VPK with lower skills grow faster. The 2025-26 cluster analysis reveals exactly which children those are, what their families look like, and what shapes their participation. Together, these findings give the DEL an unusually clear roadmap for action.

About This Research

The 2024-25 Ecological Study (Prior Year)

Published in June 2025 by the ECPRG at the University of Florida, this study analyzed 93,584 children enrolled in Florida VPK during the 2022-23 school year. Using advanced machine learning methods, the research identified three overarching findings:

- Children who enter VPK with lower SR scores grow significantly faster during the program year, a pattern called compensatory growth.
- Attendance is the single most powerful predictor of growth across all demographic groups. Children attending more than 60 hours per month grew 23-36% faster than peers with lower attendance.
- Children from families receiving SNAP benefits showed accelerated growth when attendance was maintained. This suggests VPK is particularly effective for children in economically challenging circumstances.

The 2025-26 Cluster Analysis (Current Year)

This year's analysis used a different but complementary approach: rather than asking what predicts growth, it asked who the children are. By examining dozens of characteristics simultaneously, including prenatal health, family structure, maternal education, birth outcomes, and more, we identified 12 distinct groups of children within the VPK population.

The result is a more granular picture of the VPK population: not just statistics, but portraits of real families with real circumstances. And when we examine how those 12 groups perform, the patterns

align almost perfectly with what last year's study predicted.

The 12 Groups of VPK Children

The table below summarizes all 12 groups identified in this year's cluster analysis, along with their school readiness scores and monthly growth rates, and how each group's findings relate to last year's ecological study. The School Readiness Score (PM1) is measured at VPK entry; higher scores indicate stronger early literacy and learning skills.

Group Name	% of Children	School Readiness Score (PM1)	Monthly Growth Rate	2024-25 Findings Align?	Risk Level
Married, graduate-educated (some)	8.3%	681	14.6/mo	✓ Confirms	Low
Bachelor's degree, married	31.7%	667	15.3/mo	✓ Confirms	Low
Doctoral degree, married	2.6%	690	14.4/mo	✓ Confirms	Low
Some college, intermediate care	15.3%	652	15.8/mo	✓ Confirms	Low–Moderate
Prenatal alcohol use, newborn complications	1.7%	653	14.9/mo	New insight	Moderate
High school-educated, elevated health concerns	25.5%	637	17.2/mo	✓ + Attendance key	Moderate
Third-trimester smoking, low SES	2.1%	644	15.3/mo	✓ Confirms	Mod–High
Foreign-born fathers, unmarried, low SES	7.7%	638	16.6/mo	✓ + New context	High
Maternal labor complications, lower education	3.8%	624	17.3/mo	✓ Highest growth	High
Second-trimester smoking, multi-risk	0.1%	640	15.3/mo	New insight	High
First-trimester-only smoking, low SES	0.3%	641	16.1/mo	New insight	High
Very premature, very low birth weight	0.8%	636	16.3/mo	✓ + Confirms need	Very High

Note: School readiness scores and growth rates are from the FY 2025-26 cluster analysis. The 2024-25 ecological study examined a different cohort (FY 2022-23) using different methods; findings are compared thematically, not numerically.

What Two Years of Research Tell Us

1. The Compensatory Growth Pattern Is Real and What We Now Know About Who It Applies To

Last year's study identified a striking inverse relationship: children who start VPK with lower scores grow faster. The 2024-25 research found that children whose mothers had only an elementary education (8th grade or below) scored an average of 609.7 at VPK entry but grew 17% faster than average during the program year. Children of mothers with doctoral degrees scored 690 at entry and grew 10% below average.

This year's cluster analysis provides further evidence of this pattern. The groups with the lowest entry scores (maternal labor complications at 624, the foreign-born father group at 638, and the high school-educated health concern group at 637) are also among the highest-growth groups (17.3, 16.6, and 17.2 points per month, respectively). The groups with the highest entry scores (doctoral-degree families at 690 and graduate-educated families at 681) show the slowest growth (14.4 and 14.6 points per month).

WHY THIS MATTERS

VPK is not just serving children who are already ahead. It is actively closing gaps. The children with the most ground to make up are making the most progress. This is exactly what a high-quality program should do, and it is happening across a wide range of family backgrounds and circumstances.

2. Attendance Is the Key That Unlocks Growth

Last year's finding that attendance is the most powerful predictor of growth is directly relevant to the groups identified this year. The 2024-25 study found that among children from families with limited formal education, high attendance was associated with growth rates 23-36% above average, the strongest attendance effect found anywhere in the data.

The Attendance Equation: What the Data Shows

- High attendance (>60 hrs/month): Growth rates 23-36% above average among low-education groups
- Low attendance (<50 hrs/month): Lower performance regardless of family background
- The attendance benefit is consistent across all groups, but the gain is largest for children who start with the lowest scores

3. Economic Circumstances Do Not Predict Failure and the Need for Consistent Support

Both studies converge on an important and hopeful message about economic circumstances. The 2024-25 ecological study found that children from SNAP-receiving families showed evidence of accelerated growth, but only when attendance was maintained. This year's cluster analysis reveals that many of the highest-growth groups are also those with the most constrained economic resources.

The high school-educated health concern group, the largest group in the analysis at 25.5% of children, has elevated rates of SNAP and WIC participation and lower school readiness scores at entry. Yet they show the second-highest monthly growth rate of any group (17.2 points per month). This reflects the power of a well-designed program meeting children where they are.

The policy implication is straightforward: limited economic resources should trigger more programmatic support. The data show clearly that when children from economically challenged families attend consistently, VPK delivers.

4. New This Year: Prenatal Health Risks Define a Distinct Group Requiring Coordinated Support

The cluster analysis adds something the prior ecological study could not directly address: the role of prenatal health behaviors and birth outcomes in shaping who children are when they arrive at VPK. Three groups were defined primarily by maternal smoking during pregnancy (in the first, second, or third trimester), and one group stood out for prenatal alcohol use combined with newborn complications.

These groups are small in size (0.1% to 2.1% of the VPK population) but represent a qualitatively different kind of challenge. The risks these children carry are biological and neurological as well as circumstantial. The effects of prenatal toxic exposures can shape attention, self-regulation, and learning in lasting ways, making these children among the most likely to benefit from individualized developmental support within VPK.

For these children in particular, waiting until VPK entry to begin coordinating support means missing years of developmental opportunity. The neurological effects of prenatal exposure do not begin at age four, neither should the response. Connecting families to developmental services, home visiting programs, and early intervention resources during infancy and toddlerhood gives these children a better foundation from which to benefit when they do enter VPK. A child whose self-regulation and attention have been supported from early on will be better positioned to use the learning environment VPK provides.

This principle extends beyond the prenatal-risk groups. Across all 12 groups identified in this analysis, earlier family contact and support translate into greater readiness to leverage VPK. The families most likely to benefit from pre-VPK outreach are not a small subset. They are the majority of VPK families. Home visiting programs, pediatric developmental screening, WIC and SNAP service contacts, and

community health partnerships all represent points of contact where child development support can be introduced long before a child walks into a VPK classroom. The more prepared children are when they arrive, developmentally and in terms of their families' familiarity with what VPK offers, the more fully they can use it. VPK is most powerful as the final stage of an early support continuum, not as the first stage of one.

**NEW
INSIGHT**

Prenatal health risk groups were not explicitly identifiable in last year's data. This year's cluster analysis reveals them as a distinct population requiring coordinated support, including referrals to pediatric developmental services, individualized learning plans within VPK, and active coordination with Early Steps providers.

A Strengthened Body of Evidence

Taken together, the two years of research provide a level of confidence that neither study could achieve alone. The ecological study provided statistically rigorous, population-level findings across nearly 100,000 children. The cluster analysis adds a more focused look at groups of students and real families that makes those findings actionable for the DEL.

Points of Strong Convergence Between Both Studies

- Lower-scoring children grow faster in VPK, confirmed in both years across different cohorts and methods
- Maternal education is the strongest predictor of school readiness at VPK entry
- Economic circumstances (SNAP/WIC participation) are associated with lower entry scores but higher growth rates
- VPK produces consistent positive effects across virtually all family and background characteristics
- Attendance is the most critical modifiable factor, particularly for children who start furthest behind

Implications for the DEL

The following priorities reflect what the combined two-year evidence base indicates the DEL can act on directly. Each is grounded in specific findings from one or both studies and is designed to be implementable within existing program structures.

Priority 1: Establish Attendance as the Central Operational Focus

The 2024-25 study identified attendance as the single most important lever for improving outcomes. Attendance is not a family problem. It is a systems problem the DEL can address. Recommended actions:

- Direct Early Learning Coalitions to develop and implement individualized attendance outreach protocols for children falling below the 50-hour monthly threshold, using existing provider reporting data to trigger contact within the first month of enrollment.
- Commission a statewide survey of attendance barriers through the Early Learning Coalitions, gathering information directly from families about what prevents consistent participation, so the DEL can identify systemic patterns and allocate support accordingly.

Priority 2: Target Pre-VPK Outreach to the High School-Educated and Health Concern Groups

At 25.5% of VPK children, this group represents the largest single opportunity for population-level impact. They enter with below-average school readiness but grow at one of the highest rates during VPK. The DEL can act by:

- Working with county health departments to facilitate parent engagement, using WIC and SNAP service delivery points and reaching families during the prenatal and infant periods, before children are VPK-eligible.
- Supporting home visiting programs (such as Healthy Families Florida) to include structured early literacy and executive function components as part of their standard service model.
- Funding community-based parent engagement workshops focused on early literacy and executive function practices at home, timed to the year before VPK entry for families in this profile.

Priority 3: Create Formal Coordination Protocols for Prenatal-Risk and Preterm Children

Children in prenatal-risk and very preterm groups need more than standard VPK, and many are already connected to Early Steps before VPK entry. The DEL can formalize what is currently informal:

- Develop a structured transition protocol between Early Steps and VPK that ensures developmental service records and individualized intervention plans accompany the child into VPK enrollment rather than requiring providers to reconstruct the child's history from scratch.
- Explore the feasibility of designated VPK classrooms with enhanced staffing ratios and provider training for children with complex developmental histories, building on models used in other states with similar programs.

Priority 4: Invest in Longitudinal Data Capacity

Both studies are limited to outcomes measured during the VPK year. Whether the compensatory growth observed during VPK persists into kindergarten and beyond is the most important unanswered question in this research program. The DEL can build toward an answer by:

- Expanding data linkages between the Early Childhood Integrated Data System and Florida's K-12 student data to enable tracking of these 12 groups through at least third grade, the point at which early literacy gains typically consolidate or diverge.
- Commissioning a follow-up study using the FY 2022-23 cohort from the 2024-25 ecological study, which is now old enough to have completed second grade, to examine whether VPK attendance patterns predicted kindergarten readiness and early elementary literacy outcomes.

Priority 5: Add Executive Function Measurement to the VPK Assessment System

One of the most significant gaps in our current understanding of children's school readiness trajectories is the absence of any systematic measure of executive function (EF) in the VPK assessment system. Executive function refers to the set of cognitive skills, including working memory, cognitive flexibility, and inhibitory control, that allow children to direct their own attention, manage impulses, and sustain engagement with learning tasks. These are not simply school skills. They are the capacities that determine how effectively a child can extract learning from any environment, whether that is a conversation with a parent at home, a structured activity in a VPK classroom, or an unstructured moment on the playground. A child with stronger EF does not just learn more in a good environment. That child learns more in every environment, because EF is the internal mechanism that makes environmental opportunity usable.

This matters because research consistently shows that EF predicts literacy growth independently, and in addition to children's initial literacy skills. When two children enter VPK with the same FAST score, we expect that the child with stronger EF, or the child whose executive function develops more rapidly during VPK, will outpace the other in literacy over time. This is true even when the quality of their classrooms and home environments is held constant. Executive function is a hidden driver of the growth trajectories we are currently observing but cannot fully explain. It is also a key reason why some children seem to benefit more from high-quality VPK than others: the program provides the environment, but EF determines how fully each child can use it.

The 12 groups identified in this year's cluster analysis almost certainly differ meaningfully in executive function. Children born very prematurely, children with prenatal alcohol or tobacco exposure, and children from high-stress household environments are all known from substantial independent research to face greater EF challenges. Incorporating an EF measure into the current assessment system would allow us to see this, to track it, and to connect it to the literacy growth patterns we observe.

Closing Thoughts

Two years of research, two different methodologies, two different cohorts of children, and a remarkably consistent set of findings. Florida's VPK program is working. It is generating meaningful growth across all 12 groups of children identified in this analysis, and that growth is consistently the largest among the children who arrive with the lowest school readiness scores. This compensatory pattern, replicated across two independent cohorts and two different analytic approaches, is one of the most consistent findings in this research program.

What these studies also reveal is that the program's full potential has not yet been realized. The groups with the highest growth rates during VPK are also, consistently, the groups with the lowest school readiness scores at entry and the greatest concentration of economic, health, and family risk factors. Whether those factors directly limit attendance and participation, or whether the association reflects other unmeasured dynamics, is a question this research cannot fully answer. What the data do show is that attendance is strongly associated with growth, and that the groups with the most to gain are not always the groups attending most consistently. That gap is worth understanding better, and worth acting on even before the causal mechanisms are fully established.

And there is something important we still cannot see: the role of executive function in shaping who grows, how fast, and why. Adding that lens to this research program would give the DEL tools it does not currently have, moving beyond describing outcomes to understanding and shaping them.

Technical Appendix

Sample Description, Variable Definitions, and Analytic Methods

Sample and Variable Description

Data Source and Study Population

The analytic sample was drawn from the Florida Early Childhood Integrated Data System (ECIDS), which integrates administrative records across Florida's early learning programs, vital statistics, and social service systems. The dataset used for this analysis linked VPK program records for the 2024-25 academic year with birth certificate data, early childhood program participation records, and classroom observation data. The source file comprised records uniquely identified at the child level.

The analytic sample was restricted to children born in Florida and for whom complete data were available on both primary outcomes: school readiness score at VPK entry (PM1) and average monthly literacy growth during the VPK year (AvgGrowthPerVPKStandardMonth). Children missing either outcome were excluded prior to analysis. The resulting analytic sample reflects children enrolled in Florida VPK during the 2024-25 program year with valid pre- and post-assessment data sufficient to compute both a baseline readiness score and a growth trajectory.

Outcome Variables

Two outcomes anchored both the analytic approach and the interpretation of cluster profiles. These outcomes were measured for every child in the analytic sample and served a dual function: they were used to supervise the dimensionality reduction phase of the analysis (described in the Methods section), and they provided the primary interpretive framework for characterizing each identified group.

- **School Readiness Score at VPK Entry (PM1):** The PM1 score is derived from the Florida Assessment of Student Thinking (FAST) and reflects early literacy and foundational academic skills measured at the start of the VPK year. Higher scores indicate stronger school readiness at program entry.
- **Average Monthly Literacy Growth (AvgGrowthPerVPKStandardMonth):** This variable captures the average gain in FAST score per standardized VPK program month across the program year. It reflects learning rate during VPK participation, independently of where a child started.

Predictor and Outcome Variables

Thirty-seven predictor variables were drawn from four domains: classroom environment, maternal and family characteristics, birth and perinatal health, and program participation. These variables were organized into three measurement types for analytic purposes. Table 1 provides a complete listing.

Table 1. Predictor and Outcome Variables Included in the Analysis

Category	Variable	Role in Analysis
Outcomes	PM1 (Florida FAST school readiness score at VPK entry)	Supervised outcome; excluded from predictor matrix
	Average Growth Per VPK Standard Month (AvgGrowthPerVPKStandardMonth)	Supervised outcome; excluded from predictor matrix
Continuous Predictors	Total Hours in VPK (TotalHoursInVPK)	Predictor
	CLASS: Behavior Management	Predictor
	CLASS: Concept Development	Predictor
	CLASS: Instructional Learning Formats	Predictor
	CLASS: Language Modeling	Predictor
	CLASS: Negative Climate	Predictor
	CLASS: Positive Climate	Predictor
	CLASS: Productivity	Predictor
	CLASS: Quality of Feedback	Predictor
	CLASS: Regard for Student Perspectives	Predictor
	CLASS: Teacher Sensitivity	Predictor
	Mother's Age at Delivery (MOTHER_AGE)	Predictor
	Pre-Pregnancy BMI (PrePregnancy_BMI)	Predictor
	5-Minute APGAR Score (APGAR_5)	Predictor

Binary Predictors	School Readiness Program Participation (SR_Flag)	Predictor
	Mother Born Outside the U.S. (MotherNotBornInUS)	Predictor
	Father Born Outside the U.S. (FatherNotBornInUS)	Predictor
	Mother Married at Time of Birth (MOTHER_MARRIED)	Predictor
	Father Listed on Birth Certificate (FatherOnBirthCert)	Predictor
	Premature Birth (PrematureFlag)	Predictor
	Mother Received WIC (MOTHER_WIC_YESNO)	Predictor
	Infant Breastfed (INFANT_BREASTFED)	Predictor
	Any Maternal Risk Factors (MR_ANY)	Predictor
	Any Infant Risk Factors (INF_ANY)	Predictor
	Any Maternal Morbidity Factors (MM_ANY)	Predictor
	Any Characteristics Risk Factors (CHAR_ANY)	Predictor
	Any Abnormal Conditions at Birth (AC_ANY)	Predictor
	Any Congenital Anomalies (ANOM_ANY)	Predictor
	Prenatal Alcohol Use (ALCOHOL_USE)	Predictor
	SNAP Program Participation (SNAP_Flag)	Predictor
	Early Steps Program Participation (EarlySteps_Flag)	Predictor
Categorical Predictors	Trimester of Last Smoking (LAST_TRI_SMOKED)	Predictor
	Maternal Education Level (MOTHER_EDCODE)	Predictor
	Birth Weight Category (BIRTH_WEIGHT_CAT)	Predictor
	Kotelchuck Index of Prenatal Care Adequacy (KotelChuck_Index)	Predictor

Continuous Predictors

Fourteen continuous predictors were included. Eleven were drawn from the Classroom Assessment Scoring System (CLASS), capturing dimensions of classroom quality across emotional support, classroom organization, and instructional support domains: Behavior Management, Concept Development, Instructional Learning Formats, Language Modeling, Negative Climate, Positive Climate, Productivity, Quality of Feedback, Regard for Student Perspectives, and Teacher Sensitivity. Total hours of VPK participation (TotalHoursInVPK) captured program dosage. Two birth-related continuous measures were included: mother's age at delivery (MOTHER_AGE), pre-pregnancy body mass index (PrePregnancy_BMI), and the 5-minute APGAR score (APGAR_5), an indicator of newborn health status immediately following birth

Binary Predictors

Seventeen binary predictors captured family structure, program participation, nativity, and perinatal health indicators. Family and household variables included maternal marital status at birth, whether the father was listed on the birth certificate, and whether the mother or father was born outside the United States. Program participation indicators included SR program participation (SR_Flag), WIC receipt, SNAP participation, and Early Steps early intervention program enrollment. Infant feeding was captured through a breastfeeding indicator. Six perinatal health indicators were derived from birth certificate data: a premature birth flag, and composite flags for any maternal risk factors, any infant risk factors, any maternal morbidity conditions, any perinatal characteristics risk factors, any abnormal conditions at birth, and any congenital anomalies. Prenatal alcohol use was included as a binary indicator.

Categorical Predictors

Four categorical variables captured ordinal or nominal characteristics not reducible to binary indicators. Maternal education level (MOTHER_EDCODE) reflects the highest level of education completed by the mother, ranging from elementary school through doctoral degree. The trimester of last smoking during pregnancy (LAST_TRI_SMOKED) distinguishes no smoking, first-trimester-only smoking, second-trimester smoking, and third-trimester smoking. Birth weight category (BIRTH_WEIGHT_CAT) classifies newborns into standard clinical weight groups. The Kotelchuck Index (KotelChuck_Index) is a validated composite measure of prenatal care adequacy that accounts for both the timing of care initiation and the number of prenatal visits relative to gestational age.

Missing Data

Missing data were present in predictor variables but not in the two outcomes, which served as the basis for sample inclusion. Prior to the clustering analysis, missing predictor values were addressed using multiple imputation by chained equations (MICE), implemented using the miceforest package

with LightGBM as the imputation engine. Twenty imputed datasets were generated using 20 iterations of the MICE algorithm. Mean matching was disabled to ensure numerical stability across the full predictor space. All subsequent analyses were conducted across all 20 imputed datasets, with findings pooled using Rubin's rules for combining multiply imputed results. This approach ensures that uncertainty due to missing data is appropriately reflected in the final cluster profiles and contrast estimates.

Analytic Methods

Overview

The goal of the analysis was to identify naturally occurring groups of children within the VPK population who share similar patterns across the full set of predictor variables, and to characterize how those groups differ in their school readiness at entry and literacy growth during VPK. This required an analytic approach capable of handling high-dimensional, mixed-type predictor data across continuous, binary, and categorical variables, and capable of orienting the grouping solution toward dimensions of variation that are substantively relevant to the two outcomes of interest.

Missing predictor data were addressed prior to analysis using multiple imputation by chained equations (MICE), implemented using the `miceforest` package with LightGBM as the imputation engine. Twenty imputed datasets were generated using 20 iterations of the MICE algorithm, with mean matching disabled to ensure numerical stability across the full predictor space. All subsequent analyses were conducted on these imputed datasets, with findings pooled using Rubin's rules for combining multiply imputed results.

The grouping analysis was conducted using a proprietary, patent-pending machine learning procedure developed by the ECPRG at the University of Florida. The procedure is built around an autoencoder-based dimensionality reduction step that compresses the full predictor space into a compact latent representation, followed by a clustering step that identifies stable groupings in that latent space. The number of groups identified was targeted to a range that is large enough to capture meaningful heterogeneity in the VPK population but small enough to be interpretable and actionable for policy and program purposes.

Dimensionality Reduction

The 37 predictor variables span multiple measurement types and include both highly correlated and conceptually distinct dimensions. Direct clustering on this raw predictor space would treat all dimensions as equally relevant and would be sensitive to the arbitrary scaling and distributional properties of individual variables. To address this, the predictor space was first compressed into a lower-dimensional latent representation prior to clustering.

This compression was accomplished using an autoencoder: a neural network architecture that

learns to represent high-dimensional input data in a compact form by training the network to reconstruct the original inputs from that compact representation. The autoencoder used in this analysis is a proprietary, patent-pending design developed at the University of Florida and is specifically adapted for large-scale, mixed-type administrative data of the kind produced by early childhood program records. The resulting latent representation captures the primary dimensions of variation in children's individual, family, classroom, and program characteristics, oriented toward the dimensions most relevant to school readiness at VPK entry and literacy growth during the program year.

Cluster Identification

Groupings were identified in the latent space produced by the dimensionality reduction step. The clustering procedure is part of the same proprietary, patent-pending technology and was designed to produce solutions that are stable and replicable given the structure of the data. The final solution identified 12 groups of children. The number of groups was determined by the configuration search rather than specified in advance, within the target range established prior to analysis.

Cluster Profiling

Following identification of the final cluster solution, each cluster was characterized using pooled estimates of all predictor variables and both outcomes across the 20 imputed datasets. For each predictor, the cluster-level mean or proportion was compared to the population-level mean or proportion. Predictors were flagged as prominent for a given cluster if the absolute difference between the cluster value and the population value exceeded a threshold of 0.25 population standard deviations. For continuous predictors, the population standard deviation was the standard deviation of the variable across all children. For binary and categorical level indicators, the population standard deviation was computed as the Bernoulli standard deviation corresponding to the population proportion.

This prominence threshold was selected to identify predictors that meaningfully distinguish a cluster from the overall population without requiring extreme differences. All estimates were pooled across the 20 imputed datasets using Rubin's rules, and 95% confidence intervals were computed for all cluster-population contrasts. Each group is characterized by the combination of predictors for which it differs meaningfully from the overall VPK population, and by its pooled mean scores on both outcomes.

Computational Resources

All analyses were conducted on the University of Florida's HiPerGator supercomputing system, which provides GPU-accelerated processing, enabling the training of neural network models on large administrative datasets within feasible timeframes. Imputation and profiling computations were conducted in Python using open-source scientific computing libraries.