Art. #2654, 10 pages, https://doi.org/10.15700/saje.v46n2a2654

Statistical mapping of factor relationships in the South African public primary school system

Hannah Boëttger (née Quass) and Lieschen Venter
Department of Logistics, Stellenbosch University, Stellenbosch, South Africa
lventer@sun.ac.za

Abstract
In this paper we examine socio-economic and household factors linked to learner progression in South African public primary schools, focusing on the Western Cape and Gauteng provinces. We analysed the 2022 General Household Survey using factor analysis, multilevel regression, and structural equation modelling to develop a statistically valid factor map of the public primary school system. Predictive models validated these relationships and revealed key socio-economic influences, highlighting the importance of household stability, parental involvement, and early childhood conditions for academic success. The findings show that family wealth, household stability, and school-based support were the strongest predictors of learner progression, clarifying how conditions outside the school environment shaped outcomes and offering insights for policy. The study provides a foundation for future research into targeted interventions to reduce educational inequality in South Africa.

Keywords: factor mapping; feature analysis; policy setting; public primary school system; statistical analysis

Introduction
The South African Department of Basic Education (DBE, 2024) oversees primary and secondary education, including early childhood development and special needs education. Learners progress through four phases: The Foundation Phase (Grade R to 3), the Intermediate Phase (Grade 4 to 6), the Senior Phase (Grade 7 to 9), and the Further Education And Training Phase (Grade 10 to 12), culminating in the National Senior Certificate (NSC) examination. Attendance is mandatory up to Grade 9, after which learners can opt to pursue the NSC.

In 2022, 94.5% of the 13.4 million South African learners were enrolled in public schools, with the remainder attended independent schools. Public schools are divided into quintiles based on socio-economic factors. Quintile 1 to 3 schools are no-fee schools, serving lower-income learners and receiving government funding. In 2019, 87% of public schools were no-fee schools, accommodating 79% of South African learners. Quintile 4 and 5 schools, attended by learners from wealthier socio-economic backgrounds, charge fees ranging from R2,000 to over R20,000 annually.

Education in South Africa is heavily influenced by persistent economic inequality, shown by a Gini score of 63 in 2023, which indicates one of the highest levels of income inequality globally and reflects deep disparities in access to quality education and resources. Public schools largely serve lower-income families, while private schools cater for wealthier learners, revealing disparities in resources and learning environments.

South Africa's education system faces severe challenges, consistently ranking low in international assessments. In the 2021 Progress in International Reading Literacy Study (PIRLS) reading assessment, South Africa ranked last, with many learners struggling to achieve even basic literacy and numeracy, showing that early learning outcomes mirror broader socio-economic inequality. Similarly, the 2019 results of the Trends in International Mathematics and Science Study (TIMSS) show that only 37% of Grade 5 mathematics learners and 28% of science learners demonstrated basic proficiency, illustrating the persistent gap between well-resourced and under-resourced schools. Quintile 1 to 4 schools had significantly higher failure rates in Grade 9 mathematics and science compared to Quintile 5 schools, highlighting socio-economic disparities in educational outcomes.

Van der Berg, Hoadley, Galant, Van Wyk and Bohmer (2022) found a decline in academic performance across Grades 3, 6, and 9, with language and mathematics scores in 2021 lower than in 2019. Spaull's (2013) analysis of South Africa's bimodal education system further shows that wealthier learners perform well while disadvantaged learners fall behind. These disparities stem from socio-economic inequality, as wealthier families can afford qualified teachers, smaller classes, and richer learning environments, while poorer learners attend under-resourced, underperforming schools.

Multidimensional poverty further exacerbates inequality, with 62% of South African children under the age of 17 being multidimensionally poor, facing deprivations in education, health, and housing. This deepens the educational crisis in the country.

With this study we mapped factors influencing learner progression from Grade 1 to 7 in South African public primary schools. We identified key system-level factors and their interrelationships using data from the Western Cape and Gauteng – provinces with similar yet distinct socio-economic contexts. Gauteng reflects urban challenges, while the Western Cape demonstrates relatively effective public schooling, making them useful comparative cases. Although national performance has been widely studied, little empirical work examines province-level variation in primary school progression or how household and learner factors interact across these contexts.

Literature Review

Grade repetition is a key educational measure, signalling past academic challenges and predicting future struggles. It is costly, strains resources, and indicates areas where the education system may be failing. Understanding the factors behind grade repetition can help to identify where improvements are necessary.

Wills (2023) examined early grade repetition in no-fee schools in the North West and Mpumalanga provinces, finding that learners repeating between Grades 1, 2 or 3 began school with weaker skills – low phonological awareness and minimal alphabetic literacy, scoring about 20 to 25% lower on early reading assessment than their peers. These results highlight the need for early remedial programmes and stronger Grade R provision. Similarly, Venter (2020) shows through simulation modelling that better early preparation reduced grade repetition.

Van der Berg, Wills, Selkirk, Adams and Van Wyk (2019), analysing the 2007 National Income Dynamics Study, found a strong correlation between grade repetition and school dropout rates, with home background and school quality as key factors. Learners in high-quality schools were less likely to drop out despite being behind, while racial gaps persisted: Black African learners spent nearly a year longer in school from Grade 1 to matric than White learners, and boys were 4% more likely to repeat than girls. Although national patterns of repetition and inequality are well established, little work quantifies how these dynamics interact provincially in the primary grades. We addressed this evidence gap through a comparison between the Western Cape and Gauteng.

Recent work increasingly applies systems thinking and simulation to education research, recognising that schooling outcomes emerge from complex, interdependent processes rather than isolated variables. Mason (2021) argues that education systems function as adaptive networks shaped by feedback between learners, teachers, and institutions, requiring analytical approaches that capture this complexity. Squires (2021) extends this perspective by showing how system dynamics can be used to test reform strategies and identify leverage points for improving equity and efficiency. Similarly, Yoon, Goh and Park (2022) demonstrate how systems-based modelling reveals interaction between social and institutional factors that drive educational change, offering a richer understanding of policy effects over time.

Building on this, Van den Heever, Becker, Venter and Bekker (2024) combined machine learning (ML) with agent-based simulation to model learner progression in public high schools. Using factor analysis (FA) to identify key latent constructs, they developed a framework to simulate progression from Grade 8 to 12, validated with data from a learner unit record information and tracking system (LURITS). This demonstrated the value of integrating ML and simulation to gain deeper insights into the education system.

Van Wyk (2021) conducted a longitudinal study tracking learners from Grade 1 in 2007 to Grade 12 in 2018, revealing high secondary school attrition rates but declining repetition, indicating improved internal efficiency. However, the study showed the limitations of post-hoc analysis to address system-wide interventions with simulation modelling offering a more forward-looking approach.

Gustafsson and Taylor (2022), using PIRLS data, examined improvements in South African education from 2002 to 2019, finding that home environments played a significant role in improving Grade 4 reading outcomes. Their study suggests a holistic view of education, emphasising the role of parents and communities alongside policymakers and institutions.

Similar patterns have been observed in other emerging economies. For example, Ferrão and Alves (2023) analysed grade repetition in Brazilian primary education and found that socio-demographic inequalities remained a key determinant of repetition, although the overall probability had declined over time. This comparison offers a useful international context for understanding South Africa's ongoing challenges.

Our study also strengthens the theoretical framing of educational inequality by demonstrating how quantitative models can represent complex system relationships. By linking statistical mapping to system-based perspectives on education, the analysis illustrates how structural conditions at household and school levels interact to influence learner outcomes. This integration shows how empirical modelling can extend existing frameworks that describe education as an interconnected socio-economic system.

Methodology

The General Household Survey (GHS) is an annual survey conducted by Statistics South Africa (StatsSA) to gather data on the living conditions of South African households, focusing on education, health, social development, housing, living conditions and food security. Since 2019, data collection has involved computer-assisted interviews, sampling around 20,000 households yearly, with an 88.6% response rate. The GHS dataset represents urban, tribal, and agricultural communities across provinces.

The 2022 GHS data were selected as it reflects post-COVID (coronavirus disease of 2019) normalisation. The survey covers 33,000 dwellings using a stratified two-stage design for greater precision. After cleaning, 2,284 records remained for children aged 6 to 13 years in public primary

schools in the Western Cape and Gauteng. As these provinces differ demographically from the national average, the results are regionally representative rather than nationally generalisable.

The person and household datasets were merged to link each child's record with household characteristics such as the head's marital status, education level, and income. Categorical variables were aggregated, and one-hot encoding (a method for turning categories into numerical indicators) converted nominal data into numerical indicators for analysis. This allowed inclusion of variables like school type and income without imposing order, while combining rare categories improved model stability and reduced sparse-data bias.

FA identifies groups of related variables that explain shared variance, which allowed us to isolate clusters of socio-economic and household characteristics influencing learner progression in public primary schools. FA reduces the number of variables while retaining key information. Exploratory factor analysis (EFA) was applied to the GHS data, with factor retention based on Cattell's scree test, suitable for large exploratory surveys where underlying constructs are not predefined.

Data suitability for FA were verified using the Kaiser-Meyer-Olkin (KMO) measure, with values above 0.7 deemed adequate. Variables with low KMO scores were excluded. Varimax rotation was used to improve interpretability, as low inter-factor correlations ($r < 0.3$) indicate that an orthogonal solution was more appropriate than oblique alternatives such as Promax.

After EFA, confirmatory factor analysis (CFA) was conducted to validate the hypothesised factor structure. EFA identified patterns without assumptions, while CFA confirmed the robustness of these constructs for further analysis. Model fit was assessed using the standardised root-mean-square residual (SRMR) and the comparative fit index (CFI), with values below 0.055 and above 0.94 respectively, indicating good fit. Factor loadings were then used to generate composite scores as weighted sums of indicator variables for regression analysis.

Multilevel regression analysis (MRA) extends traditional linear regression to account for hierarchically structured data, such as individual learners nested within households. This technique is particularly useful when variation at different levels (e.g., household and individual) need to be captured rather than averaged out.

Random intercept models allow the outcome intercept to vary across households while keeping predictor effects constant. These were compared with fully hierarchical models to test whether learner- and household-level relationships varied significantly across households, addressing the study's aim of quantifying multilevel influences on learner progression. Fully hierarchical models extend this by allowing both the intercept and slopes to vary, important when factors such as household wealth may differ in effect among households. Model fit was assessed using the Bayesian information criteria (BIC) and Akaike information criterion (AIC) with lower values indicating better performance. The intraclass correlation coefficient (ICC) measured variance within households; high ICC values supported the use of mixed models, which account for data nesting, reduced bias, and improved accuracy of standard errors compared to standard regression.

Structural equation modelling (SEM) tests how variables and latent factors relate within the data. It begins with CFA to build a measurement model, followed by path analysis to estimate direct and indirect relationships. This two-step process validates the measurement model before testing structural paths, reducing measurement error and improving reliability. Model fit was evaluated using standard indices of the CFI, root-mean-square error of approximation (RMSEA) and SRMR to ensure that the model represented the data well. SEM's ability to model latent constructs supports a deeper understanding of the interrelationships among factors in the GHS data.

Feature analysis ranks the importance of variables in predictive models. In our study, the target variable was grade repetition, and ML models were used to evaluate the importance of predictors. Techniques such as classification trees, random forests, and extreme gradient boosting (XGBoost) were applied to predict grade repetition. These models help to identify key factors contributing to grade repetition, offering insight for policy interventions.

Class imbalance, whereby non-repeaters greatly outnumber repeaters, was mitigated using the synthetic minority oversampling technique (SMOTE) and class-weight adjustments, which increased the weight of underrepresented cases during training. Despite these measures, the severe imbalance limited model stability and generalisability.

By integrating FA, MRA, SEM, and ML, we analysed factors influencing learner progression in South African public primary schools. The methods formed a sequential workflow: FA identified latent socio-economic and household factors; MRA quantified their relationships; SEM tested the overall structural model; and ML ranked predictors by importance, validating results through predictive performance. This combination of methods provides both a statistical foundation and practical understanding of how different household and learner factors interact to shape outcomes.

Results and Findings

FA was conducted on the person-level data from the 2022 GHS to identify latent constructs. Indicators

with the KMO measure of sampling adequacy (KMO-MSA) scores below 0.5 were excluded from the analysis, as this indicated that they were not suitable for FA. A scree plot indicated that three factors should be retained as they explained the most significant variance in the data. These factors were labelled as *person_F1*, *person_F2*, and *person_F3*.

The retained indicators are shown in Table 1. For ease of interpretation, factor loadings were multiplied by 100. Three factors were retained based on the scree plot, supported by eigenvalues above one (Kaiser criterion) and parallel analysis, confirming that a three-factor solution best captured shared variance. The first factor (*person_F1*) reflected family structure, including paternal presence and the marital status of household heads. The second (*person_F2*) represented institutional support, with indicators such as school meals and transport. The third (*person_F3*) captured general health, measured through cognitive and physical health variables. Only positively loaded variables were retained to ensure a well-identified model.

Table 1 FA sampling adequacy for characteristics at the person level

Indicator	KMO MSA	*person_F1* Family structure	*person_F2* School support	*person_F3* Learner health
fam_head_male	0.67	81*	-9	1
fam_married	0.68	81*	-9	0
fam_father_present	0.83	65*	-17	0
school_meals	0.71	-6	70*	7
school_no_fees	0.75	-10	67*	3
school_eat_at	0.74	-6	59*	3
school_walk_to	0.78	-12	52*	2
hlt_good_memory	0.67	-1	2	64*
hlt_good_mobility	0.72	-3	6	63*
hlt_good_comm	0.67	-3	1	52*
hlt_gen_health	0.79	1	7	45*
hlt_good_hearing	0.73	1	0	41*

Note. Asterisks indicate variables with factor loadings of 40 or higher on the rotated factor solution. Negative loading values indicate an inverse relationship between the variable and the underlying factor.

Cronbach's α was used to assess the internal consistency of the factors. Values above 0.7 indicate that the indicators loading onto a factor consistently measure the same latent construct. For *person_F1*, Cronbach's α was 0.83, indicating strong consistency. For *person_F2*, the value was 0.74, also acceptable. Although the overall α for *person_F3* was slightly below 0.70 at 0.68, the theoretical grouping of health indicators justified retaining all variables.

CFA was then applied to validate the factor structure. The RMSEA was 0.06, which was close to the recommended threshold of 0.055, and the SRMR was below the recommended 0.055 cutoff, indicating good fit. The CFI value of 0.94 also met the threshold for a good model fit, as per Hu and Bentler's (1999) recommendations. The GFI indicated a good fit, while the Tucker–Lewis Index (TLI) was slightly below the recommended cutoff for good fit, but still acceptable.

While the X^2 value was significant ($p < 0.0001$), which could suggest poor fit, this is expected with large sample sizes, such as the GHS dataset. Overall, the CFA results indicated that the measurement model provided an acceptably good fit to the data, confirming the validity of the identified factors.

EFA and CFA were conducted to investigate household-level factors. The EFA identified four key factors, each representing a distinct aspect of household conditions. Table 2 shows the initial sampling adequacy results, with high MSA values indicating good factor structure.

The food security factor (*hhold_F1*) measured variations on whether learners had access to regular, nutritious meals. The household wealth factor (*hhold_F2*) focused on differences in access to domestic luxuries, such as private transport and healthcare. The housing quality factor (*hhold_F3*) captured construction-related deficiencies in physical dwellings, while the sanitation factor (*hhold_F4*) assessed household exposure to air, water, and noise pollution. All factors demonstrated strong internal consistency, with Cronbach's α values above 0.8.

Table 2 FA sampling adequacy for characteristics at the household level

Indicator	KMO MSA	*hhold_F1* Food security	*hhold_F2* Family wealth	*hhold_F3* House problems	*hhold_F4* Home sanitation
food_enough	0.95	89*	10	−7	3
food_satisfied	0.94	88*	12	−8	0
food_not_hungry	0.91	86*	7	−6	7
food_good_variety	0.92	85*	11	−6	0
food_all_meals	0.95	84*	11	−4	5
food_not_worried	0.94	83*	11	−7	3
food_healthy	0.92	81*	12	−5	1
food_eat_daily	0.91	75*	4	−4	8
wlth_pvt_transport	0.8	10	80*	−1	6
wlth_own_vehicle	0.84	12	70*	−5	3
wlth_pvt_health_care	0.89	10	69*	−3	5
wlth_home_hot_water	0.93	13	58*	−11	14
wlth_many_bathrooms	0.89	5	54*	−8	11
wlth_garden_assets	0.9	7	50*	−3	−3
wlth_domestic_worker	0.9	4	48*	−2	0
wlth_many_lounges	0.88	1	41*	−5	8
house_weak_walls	0.75	−12	−11	89*	−8
house_weak_roof	0.76	−11	−11	88*	−7
house_weak_floor	0.86	−6	−11	79*	−10
san_clean_air	0.79	4	6	−3	71*
san_no_litter	0.81	4	14	−7	70*
san_clean_water	0.81	5	6	−5	67*
san_no_noise_pollution	0.83	1	6	−6	62*

Note. Asterisks indicate variables with factor loadings of 40 or higher on the rotated factor solution. Negative loading values indicate an inverse relationship between the variable and the underlying factor.

Initial CFA results showed substantial shared variance among several indicators within the same factor, suggesting that some error terms should be allowed to correlate to improve model fit. Three pairs of indicators were identified with high modification indices, indicating substantial shared variance not captured by the factor structure: hunger frequency and meal availability, food variety and healthiness, and private transport access and vehicle ownership.

Allowing these error terms to correlate improved model fit, the RMSEA decreased from 0.0727 to 0.0536, and the CFI increased from 0.9167 to 0.9553. According to Hu and Bentler (1999), RMSEA < 0.06 and CFI > 0.95 indicate good fit; the revised model therefore meets accepted thresholds and better represents the underlying factor structure while retaining theoretical consistency.

Despite these improvements, indicators related to welfare grant income were excluded from the final model due to poor loading values and high cross-loadings, indicating that they did not align well with the identified factors. This result, while unexpected, is supported by the data. Instead, the household wealth factor (*hhold_F2*), which includes indicators such as vehicle ownership and use of private healthcare, serves as a proxy for wealth derived from employment income. In South Africa, these indicators typically correspond to households in higher income quintiles, further supporting the decision to focus on *hhold_F2* as a representation of household wealth.

MRA was used to explore relationships among the latent factors in the GHS data. Each latent factor was used as the target variable in separate models to quantify its relationship with other factors. The regression coefficients represent the strength and direction of these relationships. Positive coefficients suggest that the target and predictor variables increase together, while negative coefficients indicate an inverse relationship.

At the person level, random effects were included based on Bayesian information criterion (BIC) values, with a reduction of 10 points or more indicating a substantially better-fitting model. For instance, when *person_F1* was the target, allowing the slope of *person_F2* to vary among households improved model fit (BIC = 4451.4 vs 4478.4 without the random effect). A similar approach was used for other person-level factors, although no random effects were needed when *person_F3* was the target. All person-level models included random intercepts. At the household level, simple linear regression was used, as household characteristics do not vary within households. This provided interpretable coefficients without unnecessary model complexity or hierarchical structure. Regression coefficients for each centred target variable are summarised in Table 3.

Table 3 Fixed-effects matrix for the seven factors at person and household level

	Family structure	School support	Child health	Food security	Family health	House problems	Home sanitation
Family structure	1.00	-0.03	0.00	0.05*	0.28*	0.06*	-0.05*
School support	-0.03	1.00	0.06*	0.02	-0.58*	-0.01	0.01
Child health	0.03	0.11*	1.00	-0.02*	0.01	-0.03	0.00
Food security	0.04*	-0.03	-0.03	1.00	0.17*	-0.15*	0.04
Family wealth	0.17*	-0.51*	0.02	0.12*	1.00	-0.11*	0.12*
House problems	0.07*	-0.02	-0.02	-0.15*	-0.17	1.00	-0.12*
Home sanitation	-0.04*	0.02*	-0.02	0.04	0.17*	-0.12*	1.00

Note. *$p < 0.05$.

ICC values were calculated for models centring each person-level factor as the target variable. The high ICC for *person_F1* (0.98) indicates that children in the same household share similar family structures. The moderate ICC for *person_F2* (0.66) suggests that siblings may receive different levels of institutional support, such as feeding schemes or transport. According to Koo and Li (2016), ICC values above 0.75 are strong, 0.5 to 0.75 moderate, and below 0.5 weak, placing these results in the high and moderate ranges, respectively. The low ICC for *person_F3* (0.12) shows that child health is influenced mainly by individual, not household-level, factors.

Path analysis builds on the MRA results by using the regression coefficients to establish initial relationships among factors. These relationships serve as the foundation for constructing a path model. After standardising the factor scores, the path model was refined by excluding relationships that were no longer statistically significant when considering all factors simultaneously. The model fit was evaluated through multiple indices: the RMSEA was 0.02, the SRMR was 0.0155, the CFI was 0.9922, and the adjusted goodness-of-fit index (AGFI) was 0.9908 – all indicating an excellent fit, although the RMSEA should be interpreted cautiously due to the small degrees of freedom of 11. The path coefficients provided a more nuanced understanding of the relationships among factors, capturing direct and indirect effects. The statistically significant paths are presented in Figure 1. The path coefficients represent the strength and direction of influence among factors, offering insight into how various aspects of household and individual factors interrelate within the dataset.

In the feature analysis for predicting grade repetition, the seven composite variables were used to train ML models. The grade repetition target variable indicates whether a learner was repeating a grade or not. The data were separated into a training set (80%) and a test set (20%) for model evaluation.

The classification tree model with SMOTE outperformed other ML models for predicting grade repetition, achieving a balanced accuracy of 0.58, minority-class recall of 0.23, F1-score of 0.60, and area under the receiver operating characteristic (ROC) curve of 0.64. The random forest with class weighting (balanced accuracy = 0.57) and XGBoost with SMOTE (0.58) performed similarly but with lower recall for repeaters. While SMOTE slightly reduced majority-class performance, it improved minority-class detection without adding substantial noise. The classification tree was chosen for its balanced precision-recall performance and interpretability, enabling feature ranking for predictors such as family wealth, institutional support, and sanitation. Figure 2 shows the feature importances of the composite variables.

Discussion

The MRA at the person level identified significant relationships between family structure, family wealth, and various factors related to a child's living standards and support at school. A positive coefficient ($\beta = 0.28$) between family structure and family wealth indicates that children from stable families, typically with both parents present, are more likely to come from wealthier households. These relationships are correlational rather than causal, describing associations instead of verified cause-effect mechanisms.

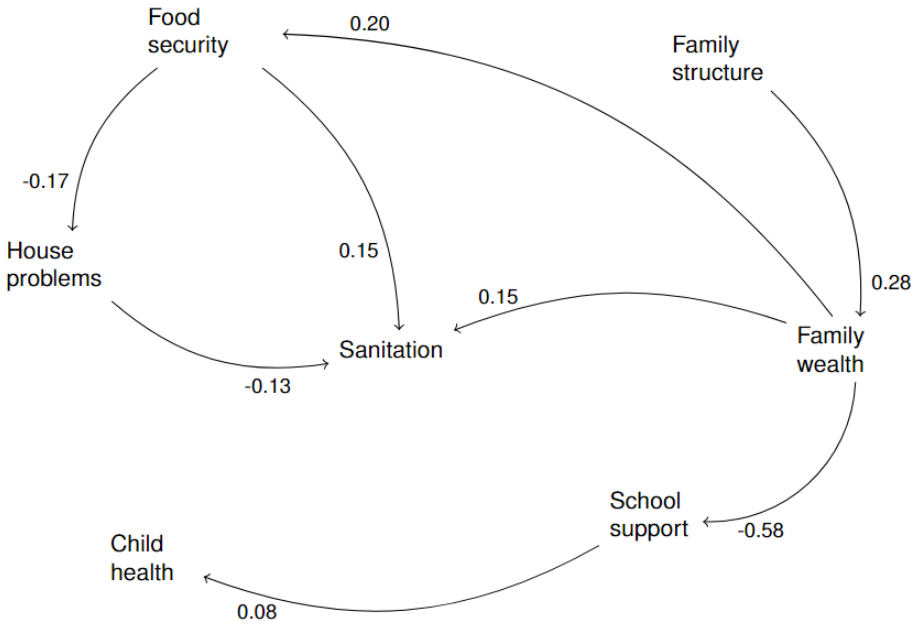

Figure 1 Relational mapping of the predictors and outcomes included in the estimated path coefficients as a network diagram

The analysis also shows a negative relationship between family wealth and institutional support received at school (β = -0.51), suggesting that children from wealthier backgrounds require less support in the form of feeding schemes or school fee exemptions. This is consistent with the observation that these children attend fee-paying schools with fewer poverty-related support mechanisms.

Further, family wealth has positive relationships with both food security (β = 0.12) and sanitation (β = 0.12), which implies that wealthier families provide their children with better access to adequate nutrition and hygienic living conditions.

The path analysis based on MRA results shows a strong association between family structure and wealth, reflecting that dual-parent households typically have higher socio-economic status. The negative correlation between family wealth and institutional support suggests that wealthier children rely less on poverty-alleviating school programmes, although this relationship is associative, not causal. A modest positive correlation between institutional support and child health ($r = 0.11$, $p < 0.05$) indicates that support such as feeding schemes benefits learners' health in less affluent households. The absence of a significant wealth-health link further implies that school-based support mainly benefits children without access to private healthcare.

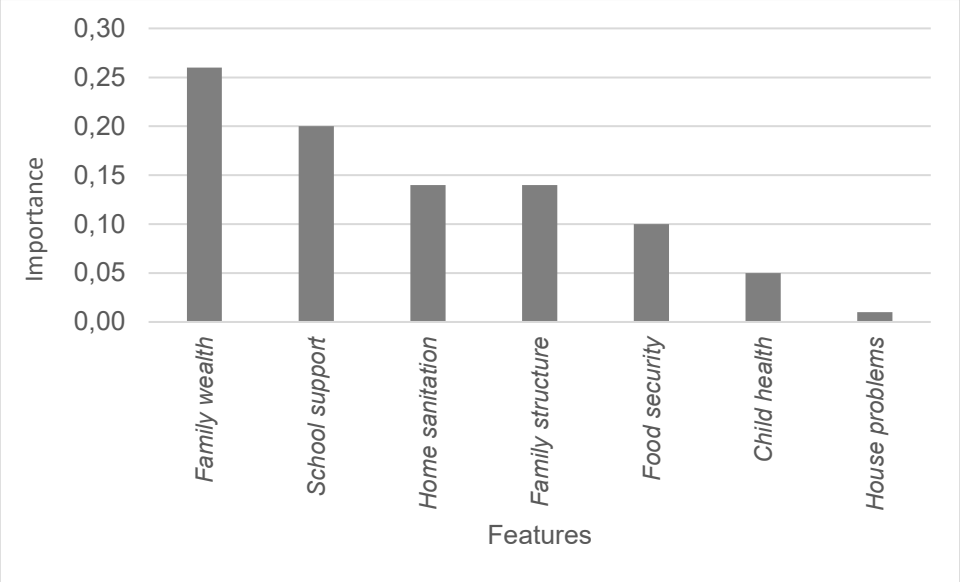

Figure 2 Feature importance of the composite variables used to predict grade repetition

Family wealth enhances household factors such as food security and sanitation and indirectly improves home construction quality through its link with food security, underscoring the interconnected nature of poverty indicators. Using model diagnostics we found no significant paths leading into family structure, supporting its classification as an exogenous variable**.** This interpretation is statistical, not causal, and the notion that altering family structure could improve well-being should be viewed cautiously given broader social and cultural factors. The findings suggest that strengthening family wealth through home-based interventions may be more effective for poverty reduction than school-based initiatives.

One limitation of this analysis is the absence of factors that measure academic success, which prevents us from quantifying how poverty affects school performance. However, the insight into home-life dynamics and poverty's multifaceted nature provides valuable context for understanding a primary school learner's life in relation to their socio-economic status.

In the relational mapping, the directional relationships between family structure and poverty are reversed compared to previous analyses, likely due to subtle differences in how family wealth and structure are defined in this study. This reversal highlights the complex causal links between family stability and wealth, especially in the South African context where these two factors are closely intertwined.

When predicting grade repetition, the most influential factor is family wealth, as confirmed by its strong relationships with institutional support, family structure, food security, and sanitation. Following family wealth, institutional support at school plays a significant role in predicting grade repetition. This suggests that while socio-economic factors like family support, sanitation, and wealth are crucial, the quality and characteristics of the school environment also matter when determining a child's academic success.

Sanitation and family structure emerge as other important factors for predicting grade repetition. These indicators rank highly in the analysis, confirming their relevance in determining whether a child repeats a grade. Child health, however, ranks as the least important factor, likely due to the low variability in health-related responses, with most children reporting no significant health challenges.

Despite the insight gained from this analysis, a key limitation is the imbalance in the dataset, with only a small percentage of the 2,284 learners (*i.e.*, 93) repeating a grade, leading to a 96% class imbalance. This imbalance causes the ML models used in the analysis to overfit the majority class, resulting in high variance in the model predictions. The instability of these predictions limits the ability to draw definitive conclusions from the ML results. High variance in this context means that the model predictions vary significantly depending on which portion of the training data is used, suggesting that the models cannot effectively generalise the factors influencing grade repetition.

Moreover, the use of grade repetition as the primary target variable for measuring educational outcomes is problematic. Given the low occurrence of grade repetition in primary school, it does not serve as an ideal measure for understanding learner progression. A better alternative would be a target variable that captures both grade progression and cognitive development, as suggested in previous studies. However, the 2022 GHS data used for this analysis lack detailed education-related information, further complicating the measurement of education

outcomes. Notably, the 2022 GHS reduced the number of education-related questions from 29 in the 2019 edition to only 15, eliminating important questions on cognitive development, literacy, numeracy, school quality, and teacher quality. This reduction in data significantly hampers education research by making it harder to identify the most influential factors that could inform interventions aimed at improving educational outcomes.

These omissions also limit the precision of the models used in this study, as several latent factors (such as cognitive ability and school environment) could not be included or validated. Consequently, relationships involving family wealth or school support may be partially overstated because learner-level performance indicators are absent. Future studies could address this by combining GHS data with administrative sources like provincial assessment datasets, or by encouraging the inclusion of cognitive and school-quality questions in future GHS rounds.

The importance of cognitive development in educational success is well documented. For example, Becker (2024) applied FA and path modelling to the 2019 GHS data, showing that cognitive development is a critical determinant of academic success at high school level. It is likely that cognitive development would be equally important for primary school learners, making its omission in the 2022 GHS data a serious limitation for studies on primary education outcomes.

Conclusion

With this study we mapped the complex relationships between socio-economic, household, and school-related factors associated with learner progression in South African public primary schools. The use of FA, MRA, and ML techniques enabled a comprehensive understanding of how family structure, household wealth, institutional support, and child health contribute to educational outcomes. The results indicate that wealthier families provide their children with better food security, sanitation, and overall living conditions, leading to more favourable academic progress. Conversely, learners from poorer households are more reliant on institutional support mechanisms such as feeding schemes, highlighting the crucial role of school support in promoting equitable education.

Addressing household-level poverty through interventions aimed at improving family wealth and stability is revealed as significant. The analysis also reveals the interconnectedness of factors like family wealth and food security, further emphasising the need for holistic approaches to address educational disparities. While school-based interventions are critical, we suggest that improving the home environment could have a more profound impact on learner outcomes.

Despite the robust insight provided, this research faced limitations. The 2022 GHS data lacks detailed measures of cognitive development, literacy, and school quality, which could further enhance the understanding of how socio-economic conditions affect learner progression. Additionally, the imbalance in grade repetition data presents challenges for ML models, limiting their predictive accuracy.

Future research should focus on integrating more comprehensive educational data to refine models of learner progression. By combining cognitive development measures with household-level socio-economic factors, future studies may provide deeper insight into the drivers of academic success in South African public primary schools.

The statistical mapping of these relationships offers a foundation for policymakers to design effective, data-driven strategies that address the root causes of educational disparities. The broader implications are interpretive and exploratory, serving to suggest possible directions for future research rather than definitive generalisations. This work contributes to efforts to understand and mitigate the influence of socio-economic inequality on learner progression, providing valuable guidance for the development of targeted educational interventions in South Africa.

These empirical limitations inform several broader implications. In practical terms, the findings suggest that policies improving household stability and economic security (such as parental employment support or conditional grants linked to school attendance) could strengthen learner progression. Expanding school-based nutrition and transport programmes in lower-quintile areas may also help to offset the disadvantages associated with household poverty. Finally, integrating education and social development data at provincial level could improve the targeting and evaluation of such interventions.

Acknowledgement

The authors used Chat Generative Pre-trained Transformer ([ChatGPT] OpenAI) to support language refinement and editorial checking. All conceptual, methodological and analytical decisions were made by the authors. The authors reviewed and verified the final manuscript and take full responsibility for its content.

Authors' Contributions

HB conducted the data preparation, statistical analysis and initial drafting of the manuscript. LV supervised the study, supported the methodological design and interpretation of the results, and revised the manuscript critically. Both authors reviewed and

approved the final manuscript.

Notes

i. Published under a Creative Commons Attribution Licence.
ii. DATES: Received: 14 October 2024; Revised: 13 October 2025; Accepted: 4 February 2026; Published: 31 May 2026.

References

Becker E 2024. *Statistical analysis for variate relationships in the South African public high school system* (Unpublished report). Stellenbosch, South Africa: Department of Logistics, Stellenbosch University.

Department of Basic Education, Republic of South Africa 2024. *Government bridges inequality gap through no fee school policy*. Available at https://www.education.gov.za/ArchivedDocuments/ArchivedArticles/Governmentbridgesinequalitygap.aspx. Accessed 5 April 2024.

Ferrão ME & Alves MTG 2023. The relationship between students' socio-demographics and the probability of grade repetition in Brazilian primary education: Is it decreasing over time? *Large-scale Assessments in Education*, 11(1):12. https://doi.org/10.1186/s40536-023-00160-0

Gustafsson M & Taylor S 2022. *What lies behind South Africa's improvements in PIRLS? An Oaxaca-Blinder analysis of the 2011 and 2016 data* (Stellenbosch Economic Working Papers: WP02/2022). Stellenbosch, South Africa: Department of Economics, University of Stellenbosch. Available at https://resep.sun.ac.za/wp-content/uploads/2022/07/wp022022.pdf. Accessed 29 May 2026.

Hu LT & Bentler PM 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55. https://doi.org/10.1080/10705519909540118

Koo TK & Li MY 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155−163. https://doi.org/10.1016/j.jcm.2016.02.012

Mason M (ed.) 2021. *Complexity theory in education futures*. Singapore: Springer.

Spaull N 2013. Poverty & privilege: Primary school inequality in South Africa. *International Journal of Educational Development*, 33(5):436–447. https://doi.org/10.1016/j.ijedudev.2012.09.009

Squires D 2021. System dynamics as a framework for education reform. *Systems*, 9(3):53.

Van den Heever M, Becker E, Venter L & Bekker JF 2024. Using machine learning and agent-based simulation to predict learner progress for the South African high school education system. *South African Journal of Industrial Engineering*, 34(3):15–27. https://doi.org//10.7166/35-3-3080

Van der Berg S, Hoadley U, Galant J, Van Wyk C & Bohmer B 2022. *Learning losses from COVID-19 in the Western Cape: Evidence from systemic tests*. Stellenbosch, South Africa: Research on Socio-Economic Policy (ReSEP), Stellenbosch University. https://doi.org/10.2139/ssrn.4212977

Van der Berg S, Wills G, Selkirk R, Adams C & Van Wyk C 2019. *The cost of repetition in South Africa* (Stellenbosch Economic Working Paper WP13/2019). Stellenbosch, South Africa: Department of Economics, Stellenbosch University.

Van Wyk C 2021. *Learner flow-through patterns in the Western Cape using CEMIS datasets from 2007 to 2019: A longitudinal cohort analysis* (Working Paper WP01/2021). Stellenbosch, South Africa: Department of Economics, Stellenbosch University. Available at https://www.ekon.sun.ac.za/wpapers/2021/wp012021/wp012021.pdf. Accessed 24 April 2024.

Venter L 2020. A systems perspective of basic education in South Africa. PhD dissertation. Stellenbosch, South Africa: Stellenbosch University. Available at https://scholar.sun.ac.za/server/api/core/bitstreams/0b732e4a-0101-4e49-bedb-2a825baabd3b/content. Accessed 15 April 2024.

Wills G 2023. *Early grade repetition*. Stellenbosch, South Africa: Research on Socio-Economic Policy (ReSEP), Department of Economics, Stellenbosch University. Available at https://resep.sun.ac.za/wp-content/uploads/2023/11/2023-11-14_Wills_early_grade_repetition_final.pdf. Accessed 15 April 2024.

Yoon SA, Goh SE & Park M 2022. Modeling the complexity of learning systems: A systems thinking approach. *Educational Research Review*, 35:100441.