Change in Early Development Metrics in Washington State Kindergarteners

Matteo Perona - Aishwarya Ramesh - Megan Pratt - Nilay Menon - Daniel Kong

University of California San Diego

I. Introduction

Statement of the Problem

Childhood development is an important determinant of health and well being later in life. Indeed, prior research has found that typical cognitive and emotional development affects their future educational, economic and health outcomes. Furthermore, child development is affected by the environment in which they grow up, including interactions with family, peers and adults around them. Much of this interaction occurs in school for young children (Likhar et al., 2022).

Development in children is also affected by socioeconomic factors such as quality of education, family income and demographic group since these affect the environment the child grows up in. Indeed, a study which examined longitudinal data collected on the developmental and health status of children from various demographic and socioeconomic backgrounds found that children of ethnic minority backgrounds were at a higher risk of negative health outcomes, as well as negative developmental outcomes such as low cognitive achievement and poor social skills. These effects varied greatly depending on which ethnic group was examined (Hillemeier et al., 2013).

Thus, in our analysis we chose to ask the question of how childhood development level has changed over time. We also aimed to examine how the trends varied with demographic group and specific global events that may have affected children such as the COVID-19 pandemic. As a case study, we chose to specifically examine the state of Washington as this state had quite a complete dataset as will be discussed in a later section.

Description of the Data

Our data comes from Washington state's open source data portal. The Washington Kindergarten Inventory of Developing Skills (or WaKIDS) dataset was collected through an observational process that was tested by the Washington Department of Early Learning during the 2010 to 2011 school year and has been used in years since. The assessment measures cognitive readiness, literacy, math, physical ability, social and emotional readiness, and combinations of each. This dataset consists of variables that measure the development levels and readiness of kindergarteners in Washington state. If the data was aggregated on the district or school level, the data will contain identifiers for the school, district, and educational service

district. The dataset also includes information about groups of students based on demographics such as gender, race, income status, and disability status. For each of these groups and levels, the data tracks kindergarten readiness in each domain either according to a "Readiness Flag" or "Development Level" where the readiness flag states whether or not a student's abilities align with kindergarten readiness, and the development level determines which age group the child's skills most align with. The measurement is listed as a percentage of the total students measured that fall into each specific group.

What Analyses have already been performed on this data?

The pilot for this project was tested on volunteer schools and classrooms in Washington State, each of which was assigned one of three observational methods for the teachers to use; Teaching Strategies GOLD (GOLD), Pearson's Work Sampling System (WSS), or CTB/McGraw Hill Developing Skills Checklist (DSC). Based on the data collected from the pilot year for this project, faculty and graduate students from the University of Washington College of Education partnered with the Washington Department of Early Learning to create a report to the Washington Legislature.

This report focused on analyzing teachers' feedback concerning which set of assessment metrics should be used in further data collection. Overall, the study determined that WSS was significantly worse according to teachers in many domains such as usefulness for planning, and timeliness. In following years, the GOLD method was used. This report only included a small discussion concerning student achievement, and this discussion was aggregated over the entire state and over all demographics. Based on analysis of students measured by the GOLD metrics, kindergarten readiness varied heavily between domains. In the social and emotional domain, close to 70% of students were meeting expectations, whereas in the physical and language domains around 45% of students were meeting expectations, and in the cognition and general knowledge domain only about 25% of students were meeting expectations. We aim to expand on this by investigating demographic disparities.

II. Preliminary Analyses (EDA)

We used EDA to explore what features we can use to help tackle our problem. Our first step was to bring the cleaned CSV files into notebooks and look through their features using pandas. With relation to our focus, *analyzing how students' learning development has changed over time*, it became clear that the most important features to explore were: School Year, Student Group, Measure, Measure Value, and Percent. The percent is the end measurement we care about within the context of the school year, since our focus relies on student outcomes over time. Next, the measure tells us which facet of education we are observing (e.g. math readiness), while the measure value was necessary in order to filter by the correct type of measurement (e.g. yes, the student is ready, or no, the student is not ready). Finally, the student groups were used to add granularity to the data; it is important to look at trends within each student group to better understand the mean patterns that emerge over time.

We chose a few visualizations which we thought were related to our problem. Figures 1 and 2 look at the 4 domains which are cognitive, literacy, math, and social emotional. Given these domains, we explored certain statistics that we thought could give us an insight towards our problem.

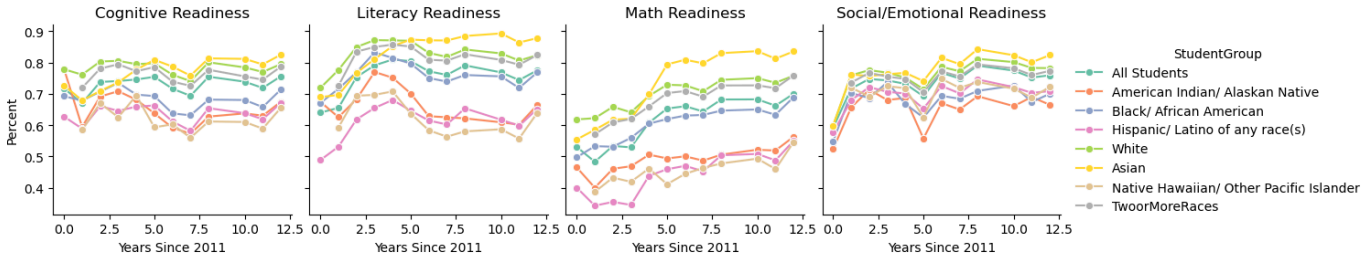

Figure 1: Change in proportion of students flagged as developmentally ready over time since 2011 in each of the 4 metrics, stratified by child's demographic group.

:

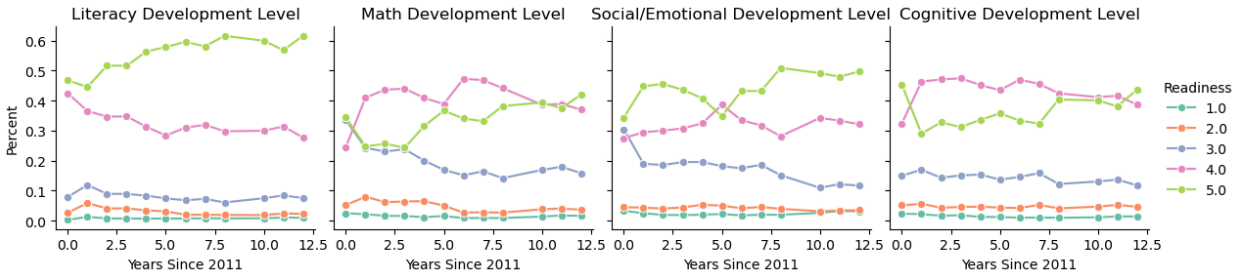

Figure 2: Change in proportion of students in each developmental level (1: 0-2 years old, 2: 2-3 years old, 3: 3-4 years old, 4: 4-5 years old, 5: kindergarten level and up) over time since 2011. Ideal development level is 5.

In Figure 1, we explored the percentage of student groups in the state of Washington that achieve a readiness level, denoting that students are passing or exceeding kindergarten level skills, from the years 2011 to 2024 across all domains. As we can see from the trends, there seems to be a decline in percentages among student groups such as Native Hawaiian, Hispanic/Latino or any race(s), and American Indian compared to other student groups across all domains. Student groups such as Asian, White, and Two or more Races seem to also be the highest in terms of percentages across the domains. This could be due to many reasons but any assumption made would have to be tested statistically.

Figure 2 again explores the 4 domains over the years 2011 to 2024, but this time focusing on the developmental levels of the students. The development stages are classified numerically with 1 representing the ages 0-2, 2 representing the ages 2-3, 3 representing the ages 3-4, 4 representing the ages 4-5, and 5 representing kindergarten and up. As we can see, the developmental levels of 1 and 2 show significantly lower percentages when compared to the

other developmental stages throughout the years. If we look across the domains, we can see that this trend seems to be constant with developmental levels 3 through 5 being higher in percentages and 1 through 2 being lower in percentages. In terms of educational developmental goals, this is favorable as the majority of students show developmental levels that are adequate for their age.

In our final section of EDA, we looked at the distributions of the percent feature in mathematical and literary readiness for all school years across different student groups to see if they exhibited significantly different shapes. The idea was that even though each student group was developing at a different base percentage level, that their rate of development could be the same or similar, and that this would be reflected in their distributions over time.

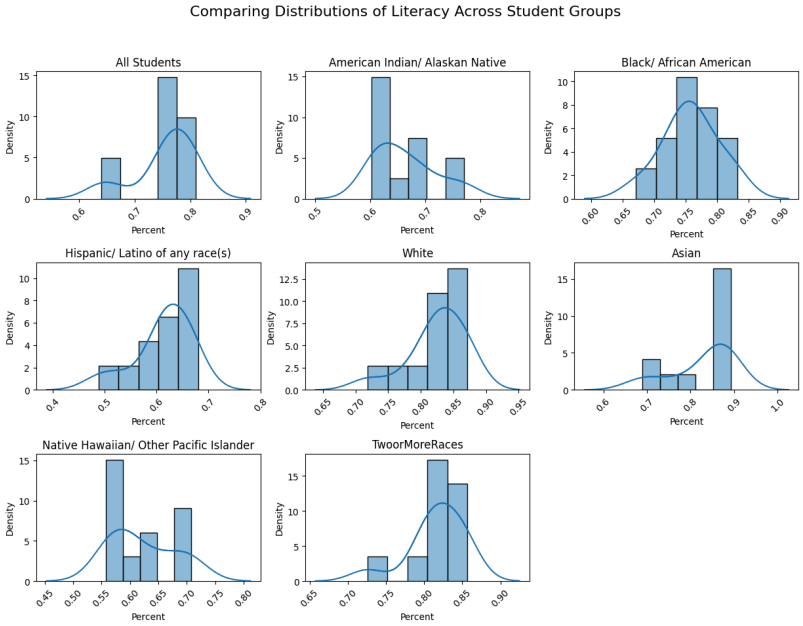

Figure 3: Distributions of the percentage of students that are deemed ready in terms of literacy across student groups. MLE Gaussian plotted over histogram to give a sense for the underlying trend.

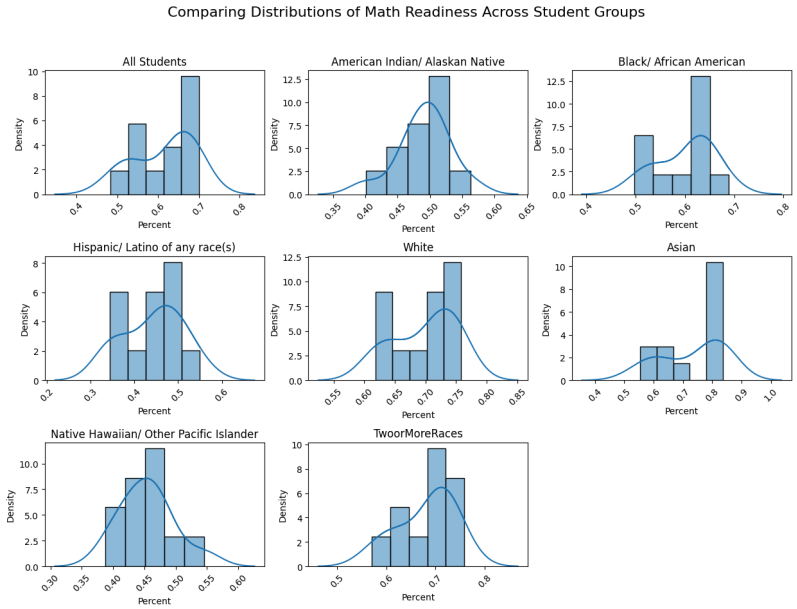

Figure 4: Distributions of the percentage of students that are deemed ready in terms of math across student groups. MLE Gaussian plotted over histogram to give a sense for the underlying trend.

We can use the number of observed means along with their placement and the magnitude of the peaks of each gaussian to illustrate patterns between plots. At a glance one can tell that most of the plots show some sort of bimodal distribution. This indicates that, over school years, percentage values across groups tend to spend more time at certain percentage values. In figure 4, measuring math readiness, most distributions tend to follow a bimodal shape where the first hump is lower than the second except for Native Hawaiian/ Other Pacific Islander and American Indian/ Alaskan native which display a more normal shape. Both, to some degree, represent populations of native people, so there could be an explanation of this similarity therein. In For literacy, figure 3, one can also see a common bimodal trend from most groups with the same outliers and one addition. This time, Native Hawaiian/ Other Pacific Islander and American Indian/ Alaskan display a bimodal distribution in the opposite direction, the second hump is smaller than the first. Notably, in literacy black student's percentages seem to come from a more normal distribution. They are the only group that shows this trend. In the results and analysis sanction we will follow up on this EDA with a Kolmogorov-Smirnov test to check which pairs of student groups' percentages come from the same distribution.

III. Methods

Data Cleaning

For the general data cleaning involved in the project, we primarily aimed to simplify the data into populations of students that we felt would be most representative of students at large. To that end, we decided to only look at groupings that included state-wide student bodies. Columns that contained data regarding location such as zip codes, school names, identification codes, and county specific tags were all removed. We then filtered the student groups to remove non-racial demographic entries such as migrant status and ESL status. Though the insights to be gained from looking at data with these groups included could yield interesting insights, they were deemed secondary to our primary goals. The data was then filtered to only include entries in specific the subject domains of cognitive function, mathematics, literacy, and social/emotional capability. We felt that these four domains provided the most representative coverage of student capability and growth metrics while maintaining simplicity in the structure of the data. After these pipelines were applied, each year's separate dataset was then checked for unexplained missing values and any entries containing incomplete data were then dropped, and combined into a comprehensive baseline to use for most specific analysis purposes.

Analyses performed for the project

In order to test our experimental question, we performed a total of 4 main analyses. First, we performed an ANOVA, using a small linear regression model using only the student group labeling by race as a predictor of the percentage of students that were marked as "ready for kindergarten" in the readiness flag for each domain. We compared this small model to a larger model that included the year that the data was collected. To ensure that our model fit the assumptions required for a regression, we tested our residuals using a QQ-Plot to check for normality, a plot of the ACF function to check independence, and a plot of the residuals against the observation to check heteroscedasticity. The residuals fit the assumptions relatively well, so we can conclude that our regression model is justified.

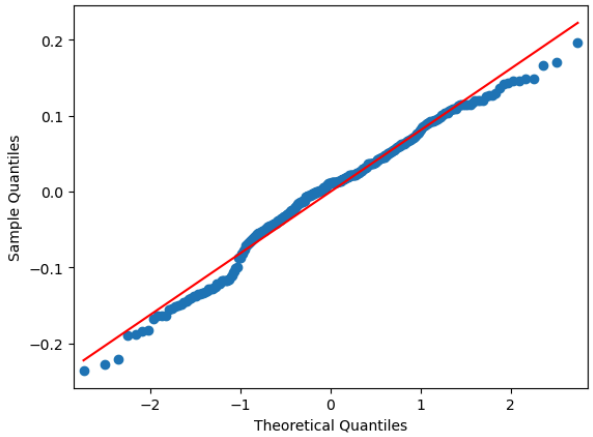

Figure 5: QQ plot of residuals compared to the normal line (in red)

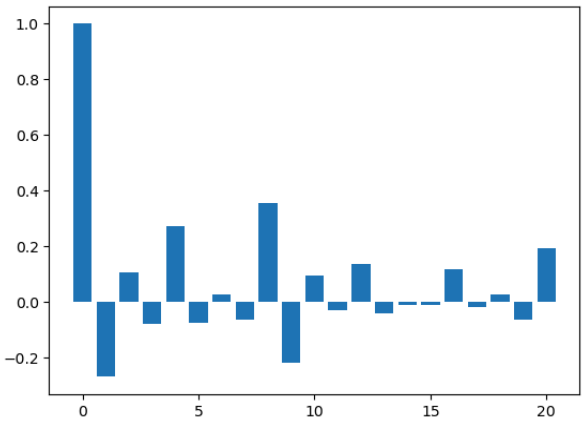

Figure 6: Plot of the ACF function of the residuals

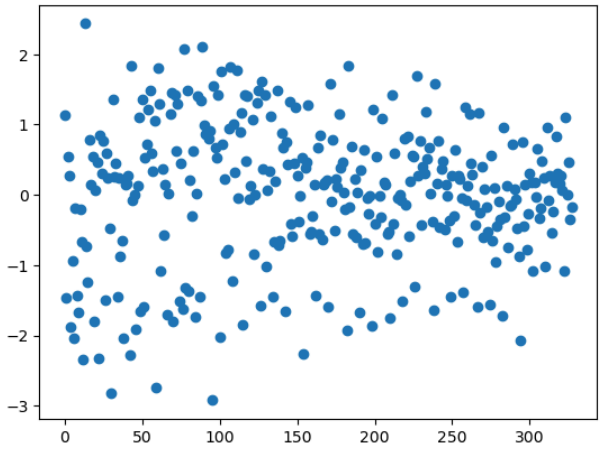

Figure 7: Plot of residuals against the observation

The second aspect of our analysis was determining whether the change in development metrics over time for different demographic groups came from the same distribution. In order to answer this question, we performed multiple 2-sample Kolmogorov-Smirnov tests comparing every demographic group with every other demographic group, and the composite student group of All Students (simply a weighted average of all separate student groups). This was done using the ks_2samp function from the scipy.stats module.

The third aspect of our analysis was the chi-square test for homogeneity. This was done to answer the question of whether the distribution of proportions of students flagged as developmentally ready in each year were from the same categorical distribution for each year. In performing the analysis, the data was first put into a contingency table where each row represented a year and each column represented a flag. The chi2_contingency function from the scipy.stats module was then used to conduct the analysis.

The final aspect of our analysis was multiple pairwise 2-sample z-tests. This was done in order to test whether the proportion of children who were flagged as being developmentally ready for kindergarten had changed significantly from year to year. This test was performed for proportions of All Students, rather than comparing specific demographic groups. The function

used to perform these analyses was proportions_ztest from the statsmodels.stats module. These results were represented in a heatmap using Seaborn. In order to correct for multiple pairwise comparisons, we employed Bonferroni correction. We conducted a total of 55 unique pairwise comparisons, and we wanted to achieve $\alpha = 0.05$. So, in order to be considered a significant relationship, each of the pairwise comparisons had to achieve a p-value of 0.05 / 55 = 0.000909. In order to accurately visualize this, the maximum value for the p-value heatmap was set to 0.000909.

IV. Results of Analyses

ANOVA

Since the plots of the residuals adhere to the assumptions of the test, we are justified in analyzing our data using the ANOVA test and linear regressions. Overall the model including school year and race as covariates only moderately fits the data, with an R^2 value of about 0.48, but we found that including the year in our model significantly increased the fit to the data.

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	321.0	2.268132	0.0	NaN	NaN	NaN
1	320.0	2.126106	1.0	0.142026	21.376345	0.000005

Table 1: Results of ANOVA test

Kolmogorov-Smirnov

We used Kolmogorov-Smirnov tests for every combination of student groups across every measurement that involves a readiness flag. Each cell shows a p-value, every cell that is cream represents a pair of student groups whose p-value passes the significance level $\alpha = 0.05$ and therefore, come from the same distribution. All of the darker cells fail to reject the null hypothesis that they come from the different distributions.

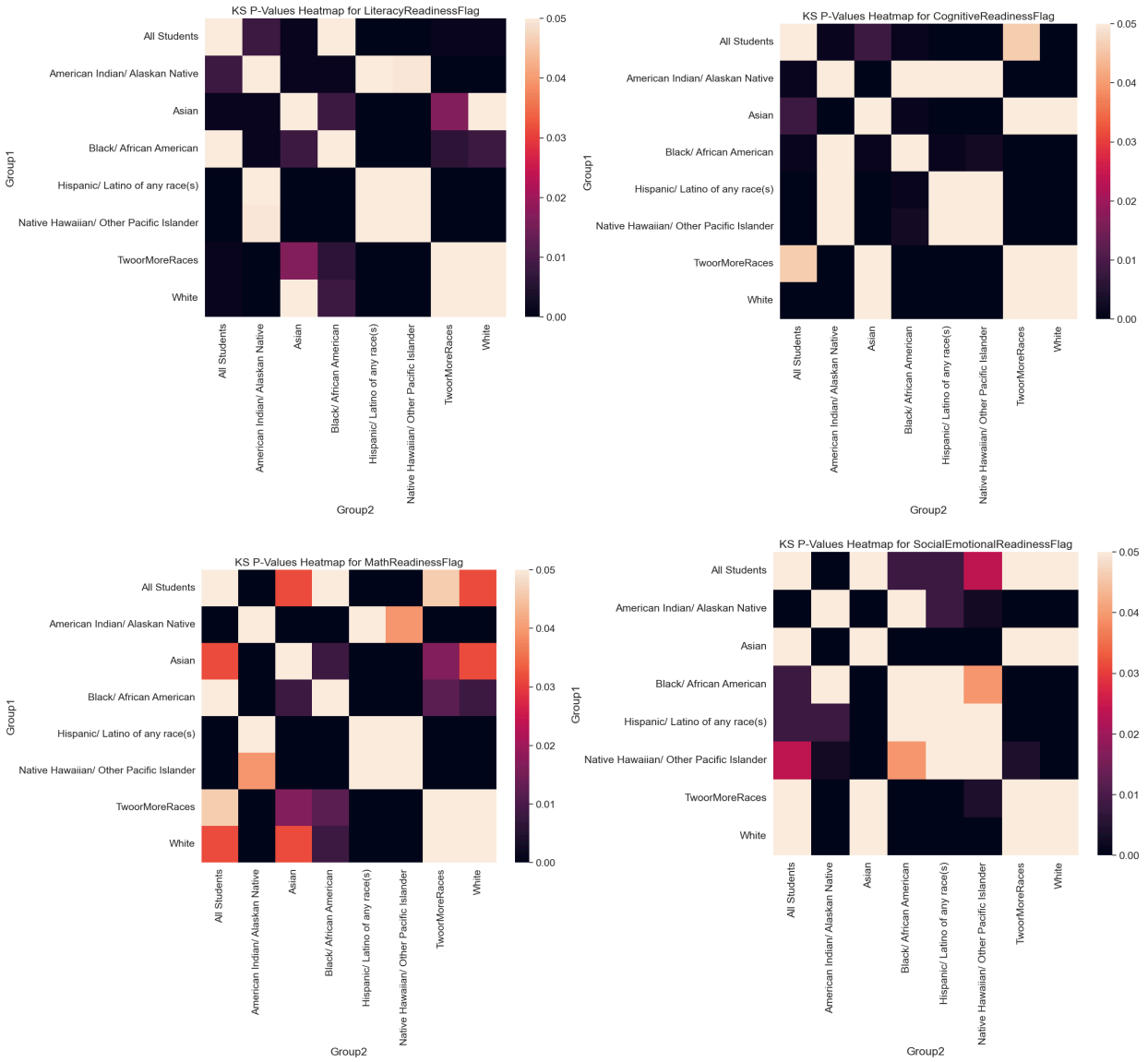

Figure 8: P-Value Heatmaps for LiteracyReadiness, CognitiveReadiness, MathReadiness, and SocialEmotionalReadiness.

As one can clearly see the diagonals all show that all pairs of student groups with themselves come from the same distribution. This gives confidence that the KS-test is working properly. For the literacy readiness plot in the top left we can see notably that Asian, White, and Two or More Races all seem to come from the same distribution. This relationship holds for all other plots except math readiness where asian students drop out of the pattern. Another notable connection is that Hispanic/ Latino of any race(s), Native Hawaiian/ Other Pacific Islander, and American Indian/ Alaskan Native seem to come from the same distribution as across all measurements except for social emotional readiness where Hispanic students drop out of the pattern.

Chi-square test for homogeneity

The application of the chi-squared test for homogeneity returned statistically significant results, indicating large differences in the distribution of student populations across the years represented in the data. The obtained p-value was extremely low with a rounded value of 0.00, leading us to reject the null hypothesis assumption of homogeneity in the distribution.

Measure	CognitiveReadinessFlag	LiteracyReadinessFlag	MathReadinessFlag	\
num_year				
0	4778.0	4263.0	3534.0	
1	14611.0	14291.0	10539.0	
2	28365.0	28966.0	20515.0	
3	31087.0	32272.0	22064.0	
4	43227.0	46159.0	35281.0	
5	58294.0	62154.0	50472.0	
6	56660.0	61108.0	52219.0	
7	55502.0	60778.0	51493.0	
8	60730.0	62714.0	54854.0	
9	55953.0	58217.0	51781.0	
10	53875.0	55715.0	49629.0	
11	56150.0	57444.0	51976.0	

Measure	SocialEmotionalReadinessFlag
num_year	
0	3940.0
1	15631.0
2	28732.0
3	31487.0
4	42642.0
5	53512.0
6	60781.0
7	60027.0
8	63688.0
9	58488.0
10	56455.0
11	56503.0

Figure 9: Contingency table where each row is a year since 2011 (not including 2020-2021) and each column is a readiness flag.

We tested the observed values of readiness flags for each subject domain in each year against the contingency table of expected observed values, given the assumption that the data from each year arose from the same distributions. The reported results (with methodology and specifics provided in the notebook) provides robust evidence against the null hypothesis, supporting an interpretation that the yearly distribution of teacher reported readiness in the four subject domains studied arise from different distributions - there are unaccounted biases from year to year that actively influence the metrics seen in our data.

The implications of these hidden variables underscore the need for a more nuanced and in-depth understanding of the distinct characteristics or factors influencing the observed variations in the data. Further exploration into population features such as student demographics or school-specific funding could yield important insights into the relationships behind the observed differences between each year.

Multiple Pairwise Z-tests

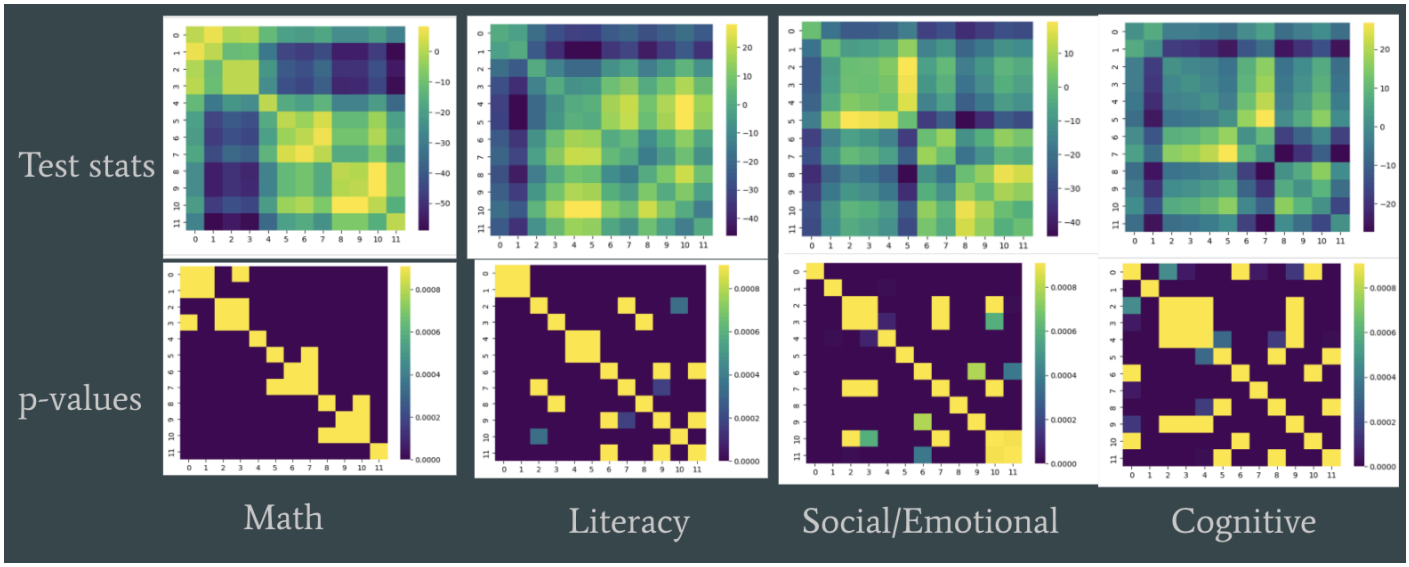

Figure 10: Axes represent years from 2011, excluding 2020-2021 since there was no data for that year. The heatmaps on the upper row represent the test statistics, with more negative test statistics being darker than the diagonal of each heatmap, and indicating a more positive difference between years. The bottom row shows a heatmap of p-values, with any non-yellow cell indicating a statistically significant difference.

From the test-stat heatmaps in Figure _, we may observe that more early years such as years 0-3 have more strongly negative test statistics when compared with later years, indicating an increase from early to later years. From the p-values, we may observe that most of these relationships are statistically significant. Especially for Math Readiness, we may observe that almost all differences are statistically significant. However, this is less true for Cognitive Development which seems to have remained fairly similar throughout the years. Overall though, we see a statistically significant increase from first to final years in all 4 measures of development. One interesting trend we may observe is that despite generally increasing , we see negative trends from years 9-10 across the board, with year 9 representing 2019-2020 and year 10 representing 2021-2022. Between that time, the major event that occurred was the COVID-19 pandemic. From the results of the z tests, we can observe the decrease in proportion of students who were developmentally ready in all 4 development metrics, with the trend being statistically significant in all but Math Readiness. While we do see statistically significant declines elsewhere as well, it is interesting to note that a significant decline occurred in ¾ metrics after the pandemic. This provides evidence for the idea that the COVID-19 pandemic may have negatively affected student development.

V. Conclusion and Discussion

Interpretation of results

From our analyses of the data, we can conclude that there is a definite difference over time, with evidence indicating an overall increase in the proportion of students who are developmentally ready since 2011. However, we note that the extent of change over time varies by demographic group. Finally, we did find evidence that COVID-19 may have affected student development negatively.

Conclusion and discussion for future work

Given the results of our investigation, we believe that future consideration of various parameters not included in our analysis, such as (but not limited to) gender, geo data, district, additional academic domains such as ESL status, extracurricular participation, and income status/neighborhood wealth would be highly valuable in determining clearer and more structured relationships of the various influences students experience and academic/growth metrics reported by their educators. There are countless studies available that link academic success in primary education and far beyond with affluence and financial circumstances, and though not studied directly here, certainly plays some part in our findings.

The data also only looks at one state, Washington, in the United States, and unfortunately does not have analogous data reports in other states with which metrics can be fairly compared as of now. While we believe there is no reason our findings should misrepresent students living in other states, we also certainly cannot claim that the findings can be extrapolated to the student population of the USA at large. Therefore, we believe that using more comprehensive statewide data with consistent metrics, whether it be through standardized exams, college acceptance rates, or future post-graduation salaries, would reveal valuable insights, which have significant potential to be considered representative of the student population across the country. We believe there is also clear value in attempting a similar study on worldwide student populations, as various areas of society such as politics, culture, and even environmental differences become accessible for in-depth comparison and mathematical analysis in regards to influences on student performance. However, non-standardized reporting and education systems actively implemented in the modern world would make such a study extremely unrealistic, to the unfortunate detriment of potentially valuable insights into the state of education for future generations.

VI. Citations

Hillemeier, M. M., Lanza, S. T., Landale, N. S., & Oropesa, R. S. (2013). Measuring early childhood health and health disparities: a new approach. *Maternal and child health journal*, *17*(10), 1852–1861. https://doi.org/10.1007/s10995-012-1205-6

Joseph, G. E., Cevasco, M., Lee, T. R., & Stull, S. (2010). *WaKIDS pilot: Preliminary report* . Washington (State), Superintendent of Public Instruction. http://www.k12.wa.us/WaKIDS/pubdocs/WaKIDS_UW%202010PreliminaryReport.pdf

Likhar, A., Baghel, P., & Patil, M. (2022). Early Childhood Development and Social Determinants. *Cureus*, *14*(9), e29500. https://doi.org/10.7759/cureus.29500