Impact of Digital Infrastructure access on the American States' educational outcomes from 2013 to 2022.

A panel data approach

Ann Mary James

University of Cincinnati

ECON8015 Panel Data Econometrics

Dr. Nayoung Lee

April 23, 2024

I. INTRODUCTION

Over the last decade, digital infrastructure has been growing at an unprecedented rate because of technological advancement, governmental initiatives, and the increasing awareness that the Internet is a vital service. This expansion is transforming sectors like education, where internet connection and computer access are necessary. Attempts to equalize the digital divide have been conducted through subsidies for broadband and devices, making digital instruments more accessible and fundamentally altering educational practices. The digital education shift sped up by the COVID-19 pandemic has highlighted the need to ensure that digital access is provided and understand the effect of digital access on educational outcomes, such as graduation rates.

This study will investigate the effect of the growing digital infrastructure on the high school graduation rates of American states from 2013 to 2022. The study will analyze how internet access and computers at home have affected educational outcomes. By analyzing panel data from this decade, the study aims to estimate digital tools' impact on student success qualitatively. Considering the different socio-economic and demographic factors, the attention will be on whether and how digital access is related to higher graduation rates.

The research question being investigated can be formalized as follows:
How does access to digital infrastructure, defined as having an internet connection or a computer at home, influence the high school graduation rates in American states from 2013 to 2022?

II. LITERATURE REVIEW

Digital literacy is more than just a skill; it's a crucial gateway to better education. This is well illustrated in the study **"The Impact of Digital Skills on Educational Outcomes: Evidence from Performance Tests"** by Pagani et al. (2015), which demonstrates how internet

skills significantly boost academic performance, mainly helping students from less privileged backgrounds and those in vocational schools.

In higher education, the impact of digital tools is varied. For instance, virtual learning environments enhance the learning process, making education more effective, while social media focuses on improving knowledge sharing. These nuances are detailed in **"Examining the Impact of Digital Technologies on Students' Higher Education Outcomes: The Case of the Virtual Learning Environment and Social Media"** by Lacka & Wong (2019).

The digital readiness of a country also plays a crucial role, particularly during challenging times like the COVID-19 pandemic. The study **"Impact of Digital Capabilities of Countries on the Pedagogical Transitions in Business Schools"** by Pandya, Cho, & Patterson (2023) highlights how nations with advanced digital infrastructures have a smoother transition to online education, pointing out the divide between digitally advanced and developing regions.

Similarly, robust digital infrastructure in the public education sector significantly enhances teaching effectiveness. This is emphasized in **"Digital Infrastructure on Teaching Effectiveness of Public-School Teachers"** by Derder, Sudaria, & Paglinawan (2023), highlighting the critical need for access to digital tools and training in digital pedagogies to elevate teaching standards.

Integrating digital infrastructure into educational environments is more than just a technological upgrade—it is a vital enhancer of learning outcomes. This underscores the importance of strategic digital planning in educational policies, ensuring that environments and capabilities are tailored to meet educational demands effectively.

III. ECONOMETRICS MODEL

All the estimators are used for panel data estimation to assess the impact of the digital infrastructure on educational outcomes in the American states from 2013 to 2022. The variable description is as follows:

Variable Description	
gradrte	Percentage of Public High School Graduates (2013 - 2022)
internet	Percentage of Households with Computer and Internet Access (2013-2022)
stutea	Student-teacher ratio (2013-2022)
linc	Log of Median Household Income from 2013-2022 (in thousands)

Table 1. Variable Descriptions

The log of median household income was taken to maintain the scale among all the variables taken for the estimation.

The model can be represented as:

$$gradrte_{it} = internet_{it} + stutea_{it} + linc_{it} + \alpha_{it} + D_t + \varepsilon_{it}$$

where $gradrte_{it}$ is the graduation rate of state i at time t.

$internet_{it}$ is the access to an internet connection and computer of state i at time t.

$stutea_{it}$ is the student-teacher ratio of state i at time t.

$linc_{it}$ is the log of median household income of state i at time t.

α_{it} is the unobserved heterogeneity.

D_{it} is the time dummy variable.

The log of median household income was taken to maintain the scale among all the variables taken for the estimation.

The time dummy variable can be defined as:

$$D_t = \delta_1 D_{2014} + \delta_2 D_{2015} + \delta_3 D_{2016} + \delta_4 D_{2017} + \delta_5 D_{2018} + \delta_6 D_{2019} + \delta_7 D_{2020} + \delta_8 D_{2021} + \delta_9 D_{2022}$$

The estimation for all the models, Pooled OLS, Fixed Effect estimator, and First Differenced estimator, was done. Additionally, Anderson and Hsiao and Arellano and Bond estimator were estimated using appropriate instrumental variables.

Several diagnostic and serial correlation tests were done to determine the best model and check the endogeneity in the model. This econometric framework will enable a comprehensive analysis of the impact of digital infrastructure on the educational outcomes of the American states from 2013-2022.

IV. DATA

The data for this study comprises panel data for all 50 U.S. states from 2013 to 2022. It includes yearly measurements of all the variables. The dependent variable is chosen as "Public High School Graduation Rate" to show the educational outcomes of the states. The independent variable is selected as the "Percentage of Households with Computer and Internet Access" to show the digital infrastructure. The control variables were chosen as the "Student-teacher ratio" and "Median Household Income." The sources for the data are as follows:

- **National Centre for Education Statistics (NCES):** NCES was used to collect data on the public high school graduation rate, percentage of households with computer and internet access, student-teacher ratio, and median household income.

The summary statistics of the variables are given below:

Summary Statistics				
Variable	Mean	Standard Deviation	Minimum	Maximum
gradrte	85.37804435	4.755571789	68.5	96.1
internet	84.12374	6.439125818	58.68	93.8
stutea	15.35875252	2.855845881	10.3	24.3
inc	63577.92	11870.89476	37963	95800
Number of observations: 500				

Table 2: Summary Statistics

The variables gradrte and internet are treated as percentages, the variable stutea is treated as a ratio, and the variable inc was transformed to linc by taking the logarithm of the median household income. A comprehensive overview of the variables used in the study is given to provide an outline.

V. RESULTS

For the estimation of the model, firstly, Pooled OLS, First Differenced, and Fixed Effect estimators were estimated. The estimation table is given as below:

Pooled OLS, FD, FE

	Dependent variable:		
	gradrte		
	OLS	panel	
		linear	
	(1)	(2)	(3)
internet	-0.023	0.043	-0.062
	(0.093)	(0.083)	(0.081)
stutea	-0.624***	0.323	0.052
	(0.066)	(0.222)	(0.187)
linc	2.309	2.148	1.379
	(2.029)	(2.452)	(2.764)
Constant	68.790***	-0.855	
	(16.921)	(0.683)	
Year Dummies	Yes	Yes	Yes
Observations	493	443	493
R2	0.360	0.379	0.537
Note:	*p<0.1; **p<0.05; ***p<0.01		

Figure 1: Estimation Results for Pooled OLS, FD, and FE

The table shows that only the variable stutea (student-teacher ratio) is significant, but only for Pooled OLS. The other variables are not statistically significant at the 5% level for the other models. The original number of observations was 500 observations, but due to the missing values, the number of observations is 493 for Pooled OLS and Fixed Effect estimator, but since

the first difference is taken for the First Differenced Estimator, the number of observations decreases for the second model that is the FD estimator. Specific tests will be carried out to check which estimator is best among these models.

The first test is to determine if the unobserved heterogeneity is present in the model or not. The null hypothesis for the test to determine the presence of unobserved heterogeneity is:

$$H_0: \sigma_\alpha^2 = 0$$

The first test to determine the presence of the unobserved heterogeneity is the Breusch Pagan test. The test result is given below:

```
> bptest(pols1)

        studentized Breusch-Pagan test

data:  pols1
BP = 62.833, df = 12, p-value = 6.839e-09
```

Figure 2: Breusch Pagan test for serial correlation

The p-value is 6.839e-09. At the 5% significance level, since the p-value is less than 0.05, we reject the null hypothesis. Therefore, there is a variance in the unobserved heterogeneity. Hence, we cannot consider Pooled OLS to be the best model. The next test to determine the presence of unobserved heterogeneity is the Serial Correlation test. The test result is given below:

```
Call:
lm(formula = pols1.res ~ pols1.lres, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-4.237e-15 -3.750e-16 -1.930e-16  3.000e-17  8.769e-14

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) -1.346e-31  1.795e-16 0.000e+00        1
pols1.lres   1.000e+00  4.713e-17 2.122e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.985e-15 on 491 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 4.501e+32 on 1 and 491 DF,  p-value: < 2.2e-16
```

Figure 3: Serial Correlation test.

The p-value is <2e-16. At the 5% significance level, since the p-value is less than 0.05, we reject the null hypothesis. Therefore, there is a variance in the unobserved heterogeneity. Hence, we drop the Pooled OLS as the best model.

Since there is an unobserved heterogeneity, the choice must be made between the Fixed Effect estimator and the First Differenced estimator. A serial correlation test, FE vs FD, determines the best model. The null hypothesis for this test is given below:

$$H_0: \rho = 0 \text{ for } \Delta \varepsilon_{it}' = \rho \Delta \varepsilon_{it-1}' + \kappa_{it}$$

The test result is given below:

```
        Wooldridge's first-difference test for serial correlation in panels

data:  fd
F = 6.4257, df1 = 1, df2 = 391, p-value = 0.01164
alternative hypothesis: serial correlation in differenced errors
```

Figure 4: FE V/s FD.

The p-value is 0.01164. At the 5% significance level, since the p-value is less than 0.05, we reject the null hypothesis. Hence, there is a serial correlation among the differenced errors. Therefore, the Fixed Effect estimator is preferred.

A Serial correlation for the level error needs to be carried out to choose FE as the best model. The null hypothesis for the test is given below:

$$\varepsilon_{it} = \delta_0 + \delta_1 \varepsilon_{it-1} + \text{error}$$

$$H_0: \delta_1 = -\frac{1}{T-1}$$

The test result is given below:

```
        Wooldridge's test for serial correlation in FE panels

data:  fe
F = 74.701, df1 = 1, df2 = 441, p-value < 2.2e-16
alternative hypothesis: serial correlation
```

Figure 5: FE Serial Correlation

The p-value is <2.2e-16. At the 5% significance level, since the p-value is less than 0.05, we reject the null hypothesis. Hence, there is a serial correlation among the level errors. Therefore, robust standard error needs to be reported. The serial correlation robust standard errors by Arellano are given below:

```
t test of coefficients:

                 Estimate Std. Error t value Pr(>|t|)
internet         -0.062481   0.159662 -0.3913 0.695742
stutea            0.052347   0.379096  0.1381 0.890240
linc              1.378539   4.758471  0.2897 0.772184
factor(year)2014  0.655350   0.513255  1.2769 0.202342
factor(year)2015  2.164637   1.937399  1.1173 0.264493
factor(year)2016  2.541566   1.264708  2.0096 0.045096 *
factor(year)2017  3.433279   2.074660  1.6549 0.098680 .
factor(year)2018  3.560689   1.822704  1.9535 0.051404 .
factor(year)2019  4.089846   1.875530  2.1806 0.029751 *
factor(year)2020  3.831857   2.309870  1.6589 0.097862 .
factor(year)2021  3.869792   2.522794  1.5339 0.125780
factor(year)2022  9.089660   2.781941  3.2674 0.001172 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Robust Standard Errors

Since none of the estimators produced statistically significant estimates, Anderson and Hsiao's estimator was estimated. The model can be written as follows:

$$\Delta gradrte_{it} = \Delta\gamma gradrte_{it-1} + \Delta internet_{it} + \Delta stutea_{it} + \Delta linc_{it} + \alpha_{it} + D_t + \varepsilon_{it}$$

where $gradrte_{it-1}$ is the lag of the dependent variable.

The estimation result is given below:

```
Call:
ivreg(formula = D.gradrte ~ DL.gradrte + D.internet + D.stutea +
    D.linc + factor(year) - 1 | L2.gradrte + D.internet + D.stutea +
    D.linc + factor(year) - 1, data = data.panel)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4331  -1.2416   0.1848   1.2723   9.6184

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
DL.gradrte        1.4188796  0.3448006   4.115 4.77e-05 ***
D.internet        0.1947342  0.1343210   1.450  0.14796
D.stutea          0.6056418  0.3736088   1.621  0.10585
D.linc            4.4744501  3.8808284   1.153  0.24966
factor(year)2015 -2.4706083  1.6721636  -1.477  0.14039
factor(year)2016  0.4733382  0.8325889   0.569  0.57003
factor(year)2017 -1.5834975  0.9349546  -1.694  0.09116 .
factor(year)2018 -0.2148132  0.5322988  -0.404  0.68677
factor(year)2019  0.0006083  0.4970041   0.001  0.99902
factor(year)2020 -1.9231753  0.7321940  -2.627  0.00898 **
factor(year)2021  0.2554615  0.5276699   0.484  0.62858
factor(year)2022  4.8531804  0.5717445   8.488 5.02e-16 ***

Diagnostic tests:
                df1 df2 statistic  p-value
Weak instruments   1 373     47.90 1.97e-11 ***
Wu-Hausman         1 372     69.87 1.28e-15 ***
Sargan             0  NA        NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.2 on 373 degrees of freedom
Multiple R-Squared: -0.2069,    Adjusted R-squared: -0.2457
Wald test: 12.85 on 12 and 373 DF,  p-value: < 2.2e-16
```

Figure 7: Anderson and Hsiao Estimator

For this model, an appropriate Instrumental variable was used, which is the $gradrte_{it-2}$. From the diagnostic tests, we can see that for the Weak Instrument test, the p-value is 1.97e-11. At the 5% significance level, since the p-value is less than 0.05, we reject the null hypothesis of Weak IV. Therefore, the model doesn't have a weak IV. For the Wu-Hausman test, the p-value is 1.28e-15. At the 5% significance level, since the p-value is less than 0.05, we reject the null hypothesis of no endogeneity. Therefore, the model has endogeneity.

Since the model consists of 10 time periods, the Anderson and Hsiao estimator produces consistent estimates but not efficient estimators. Therefore, there is a need to estimate Arellano and Bond Estimator. The model used an appropriate instrumental variable. The model can be written as follows:

$$\Delta gradrte_{it} = \Delta\gamma gradrte_{it-1} + \Delta internet_{it} + \Delta stutea_{it} + \Delta linc_{it} + \alpha_{it} + D_t + \varepsilon_{it}$$

The estimation result is given below:

```
Twoways effects Two-steps model Difference GMM

Call:
pgmm(formula = gradrte ~ lag(gradrte) + internet + stutea + linc |
    lag(gradrte, 2:99), data = data.panel, effect = "twoways",
    model = "twostep")

Unbalanced Panel: n = 50, T = 4-8, N = 385

Number of Observations Used: 284
Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-11.44761  -0.97769   0.00000  -0.04164   1.06355   9.46265

Coefficients:
             Estimate Std. Error z-value  Pr(>|z|)
lag(gradrte)  0.81038    0.19557  4.1437 3.418e-05 ***
internet     -0.14788    0.24452 -0.6048   0.54532
stutea        0.42565    0.55979  0.7604   0.44703
linc          6.06196    7.04896  0.8600   0.38980
2017          0.50829    1.39930  0.3632   0.71642
2018          0.20458    0.88296  0.2317   0.81677
2019          0.33176    1.11186  0.2984   0.76541
2020         -0.31094    1.98271 -0.1568   0.87538
2021          0.15760    2.39744  0.0657   0.94759
2022          5.65381    2.32162  2.4353   0.01488 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sargan test: chisq(20) = 28.9148 (p-value = 0.089451)
Autocorrelation test (1): normal = -2.457647 (p-value = 0.013985)
Autocorrelation test (2): normal = 1.230781 (p-value = 0.2184)
Wald test for coefficients: chisq(4) = 17.81918 (p-value = 0.0013387)
Wald test for time dummies: chisq(6) = 41.48951 (p-value = 2.3188e-07)
```

Figure 8: Arellano and Bond Estimator

From the estimate table, we can see that the γ is greater than 0.8 with the statistical significance at a 1% significance level, meaning there is evidence of weak IV. Except for the IV, no other variable is statistically significant.

The Sargan test is for the overidentification test with the null hypothesis that the moments condition is satisfied. For the Sargan test, the p-value is 0.089451. At the 5% significance level, since the p-value is greater than 0.05, we fail to reject the null hypothesis.

The Autocorrelation test is the test for the serial correlation among the errors. The differenced errors are supposed to be serially correlated in their first order but not correlated in their second order. From the test results, we can see that the model satisfies the condition. For AR (1), the p-value is 0.013985. At the 5% significance level, since the p-value is less than 0.05, we reject the null hypothesis that there is no first-order correlation among the differenced errors. For AR (2), the p-value is 0.2184. At the 5% significance level, since the p-value is greater than 0.05, we fail to reject the null hypothesis that there is no second-order correlation between the differenced errors.

VI. CONCLUSION

This paper investigated the impact of digital infrastructure on educational outcomes in the US states from 2013 to 2022. From the estimation of the models, the Fixed estimator was chosen as the best model with the robust standard errors reported. Since the estimates weren't statistically significant, Anderson and Hsiao's estimator was estimated. The time period for this study was 10 years, which created problems for the Anderson and Hsiao estimator as it gave consistent estimates but was not efficient. Therefore, the Arellano and Bond estimator was estimated with the model expressed as follows:

$$\Delta gradrte_{it} = \Delta\gamma gradrte_{it-1} + \Delta internet_{it} + \Delta stutea_{it} + \Delta linc_{it} + \alpha_{it} + D_t + \varepsilon_{it}$$

The diagnostic tests gave good results, but the variables were still statistically insignificant. Still, the best model chosen would be the Arellano and Bond estimator.

For further research, there is a possibility of seeing the effect on more control variables, such as the states' category, such as rural or urban states, poverty rates, etc. It can also consider the qualitative aspect of the digital infrastructure, such as the broadband speed of the internet connectivity in households. It can also consider the digital tools used explicitly by the students for their education purposes. Research can also be done to see the differences in various countries, such as a comparative study.

In conclusion, this study contributes to the ongoing research on the impact of digital infrastructure on educational outcomes. Although this study can include improvements, the study does talk about the effects and how the academic field can change concerning the digital infrastructure.

VII. REFERENCES

- Pagani, L., Argentin, G., Gui, M., & Stanca, L. (2015b). The impact of digital skills on educational outcomes: Evidence from performance tests. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2635471
- Lacka, E., & Wong, T. C. (2019). Examining the impact of digital technologies on students' higher education outcomes: The case of the virtual learning environment and social media. Studies in Higher Education, 46(8), 1621–1634. https://doi.org/10.1080/03075079.2019.1698533
- Pandya, B., Cho, B., & Patterson, L. (2023). Impact of digital capabilities of countries on the pedagogical transitions in business schools. *Global Knowledge, Memory and Communication*. https://doi.org/10.1108/gkmc-06-2023-0201

- Derder, A., Sudaria, R., & Paglinawan, J. (2023). Digital Infrastructure on teaching effectiveness of public-school teachers. *American Journal of Education and Practice*, *7*(6), 1–13. https://doi.org/10.47672/ajep.1719
- *Digest of Education Statistics, 2022*. National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education. (n.d.). https://nces.ed.gov/programs/digest/d22/tables/dt22_102.30.asp?current=yes
- *Digest of Education Statistics, 2017*. National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education. (n.d.-a). https://nces.ed.gov/programs/digest/d17/tables/dt17_702.60.asp
- *Digest of Education Statistics, 2016*. National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education. (n.d.-a). https://nces.ed.gov/programs/digest/d16/tables/dt16_219.20.asp
- *Common core of data (CCD)*. Table 2. Number of operating public schools and districts, student membership, teachers, and pupil/teacher ratio, by state or jurisdiction: School year 2020–21. (n.d.). https://nces.ed.gov/ccd/tables/202021_summary_2.asp