Exploring Education in Brazil

University of Mannheim Department of Political Science

Research Design

Filippo Panfoli 2202955

filippo.panfoli@students.uni-mannheim.de | PanfoliF |

0009-0000-7861-8036 | https://github.com/PanfoliF

Replication materials / LaTeX sources: PanfoliF/Research-Proposal-RD-NEW

Version: v0.0 | **Last updated:** January 9, 2026

Keywords: game theory

JEL: C72; D82

Prepared for: Research Design, taught by Prof. Dr. Sean Carey

January 9, 2026

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Introduction

This paper proposes a research project on educational data. The aim is establish causality between variables of interest using state of the art causal inferene techniques. The Brazilian context offers interesting sources of variation among data as well as granular observations. These are key factors in providing rubust empirical analysis.

Motivation: under many aspects there is little knowledge of what causes good educational outcomes (**<empty citation>**). Many factors concur in formig students, thus scholars have always had hard times in distinguishing chains of causes and effects. The proposed paper could provide some more evidence.

Data: such a high aim requires the best tools. The author put a lot of effort in the data collection step, many institutional websites have been surfed until the best source of data come out. The Brazilian Education Panel Databases (Huberts et al., 2025), appears the best source of observation for the proposed task. Not only it provides researchers with granular data, it also comes from a reliable and well know institution, namely IDB (Interamerican Development Bank).

Methodology: If we are to establish causality in a credible way, we need to use causal inference methodologies. The author will provide the reader with extensive axplantion of emplyed identification strategy in the methodology section (Methodology).The following lines briefly anticipate that section: The identification strategy selected for our aim is splet in two: first, due controls come along with two regressions. Secondly, a Difference in Difference application of these regressions, should isolate with more efficacy the causal effec of independent variables on dependent one. An appropriate review of the literature provides for suitable control variables.

Policy value: The following pages are intended to create value for policy makers. Putting aside cumbersome vocabulary and contex related dictions, they should inform policy making in the field of education in a simple and transparent

manner. The results of the analysis may open new ways of investing in education and shed light on wiser policy aims.

Research Question

Hypothesis

My goal is to distinguish two channels: standardized test performance (*extensive margin*) and grade progression outcomes (*intensive margin*). Comparing extensive and intensive margin helps determine which mechanisms affect the test score.

The analysis in the next sections will test the following hyposthesis:

$$Hp\ n^o1\text{: } Teacher\ quality \Uparrow \Rightarrow median\ student\ performance \Uparrow$$

$$Hp\ n^o2\text{: } Teacher\ quality \Uparrow \Rightarrow worst\ students\ performance \Uparrow$$

1 Literature Review and Historical Setting

Knowledge and insights on the historical context and institutional settings are drawn from Encyclopedia Britannica (Ball, James, et al., 2026; Ball, Schneider, et al., 2026), Glossario of Atlas Geográfico Escolar (CEON, n.d.) and Southey (2012).

These norms will serve as instruments to predict a changes in schools quality. They are expected to positively affect the outcome variables.

A preliminary analysis of the literature on education in developing countries, highlighted a study from Turmena and Bitencourt (2022). The journal article constitutes the milestone of this article and serves as main reference for the literature on education in Brazil. This paper suggests that the "Law No. 5.692/1971 (Reforma do Ensino de 1ř e 2ř Graus)" had a major effects on education. I will exploit data related to this law.

2 Data

2.1 Data Collection

All projects begin with data collection, which is a crucial step. However it takes a lot of time and effort.

In order to select the best source, many datasets have been explored and many institutional websites have been visited. Potential data sources included IPUMS ("IPUMS Online Data Analysis System", n.d.) and the Instituto Brasileiro de Geografia e Estatística ("Portal Do IBGE", 1967). Additionally, aggregated data may be retrieved from other sources[1]. Eventually, an article from Rubiane Daniele Cardoso de Almeida et al. (2023) offers panel data on some demographic aspects.

The Brazilian Education Panel Databases (Huberts et al., 2025), which covers the period from 1996 to 2015, was selected as main source of data.

[1](Instituto Brasileiro de Geografia e Estatística (IBGE), n.d.-a, n.d.-b, n.d.-c, n.d.-d, n.d.-e, n.d.-f, n.d.-g, n.d.-h, n.d.-i, n.d.-j, n.d.-k, n.d.-l, n.d.-m, n.d.-n, n.d.-o, n.d.-p, n.d.-q, n.d.-r, n.d.-s, n.d.-t, n.d.-u, n.d.-v, n.d.-w, n.d.-x, n.d.-y, n.d.-z, n.d.-aa, n.d.-ab, n.d.-ac, n.d.-ad, n.d.-ae; Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), n.d.-a, n.d.-b).

2.2 Description of Data

3 Methodology

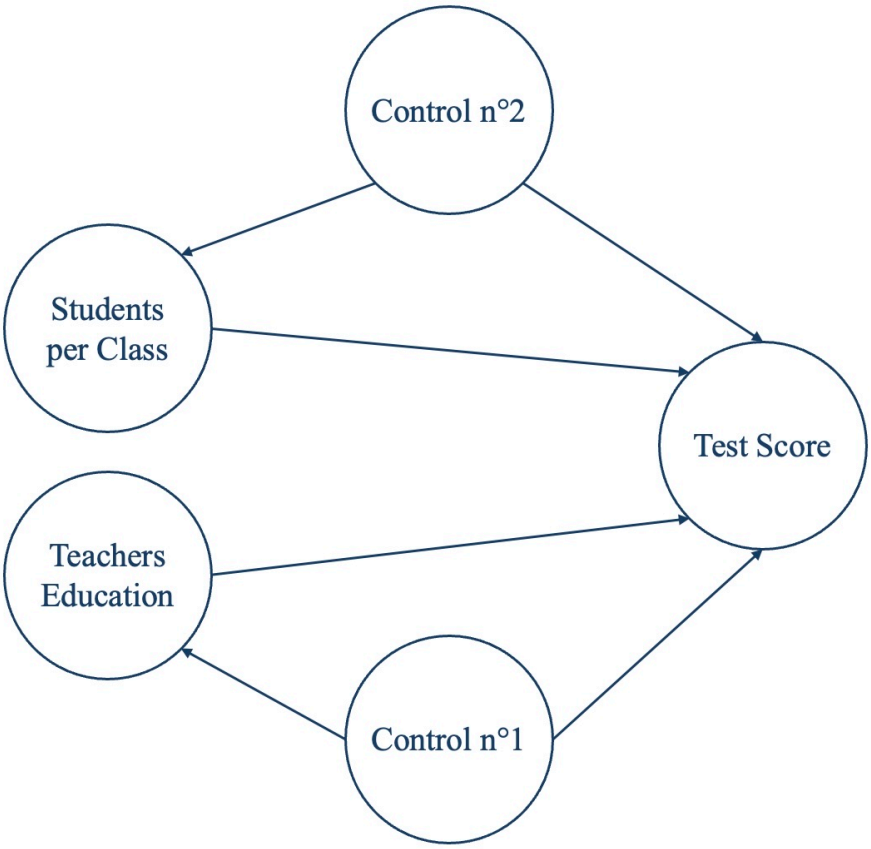

Figura 1: *Causal DAG.*

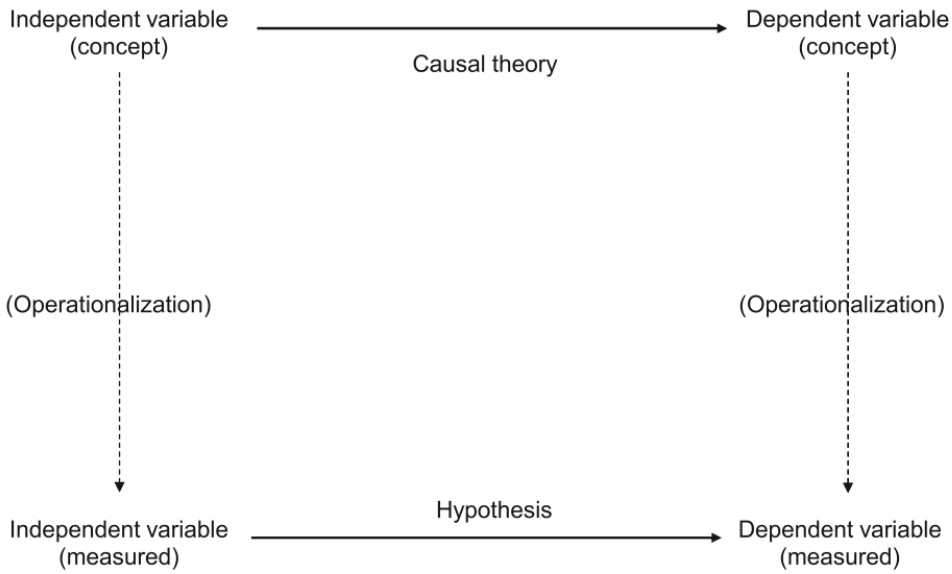

Figura 2: *Figure 1.2 From theory to hypothesis*

To establish causality between the dependent and independent variables, the analysis will employ causal inference techniques. As a possible solution to the identification problem the study will provide results from a Difference-in-Difference. This technique is able to isolate the effect produced by the introduction of the laws in the variables that approximate education quality. Nonetheless, without appropriate control variables no identification strategy is reliable. Following best practices of the political science field, only a deep analysis of the literature will provide for suitable control variables.

3.1 Extensive Margin

The first of the two regression presented in the paragraph looks at the extensive margin. Therefore, it investigates the relationship between teachers quality and test scores. The equation used is the following:

$$Y_{smt}^{score} = \beta_0 + \beta_1 TQ_{smt} + \beta_2 INFRA_{smt} + \gamma X_{smt} + \mu_m + \lambda_t + \varepsilon_{smt}$$

Teachers quality is measured as number of students per class *or* as teachers education.

Identification also relies on:

-*Municipal FE*: this is a way to control for possible unobserved variables that might bias the analysis. To the eyes of statistitians this is a mere intercept that captures the mean value for a town.

-*State FE*: the same concern we had for the municipal level, motivates the use of FE at the state level. However, concern comes along with a great fortune[2]. What Cunningham (2021, chap. 2, p. 46) said about USA perfectly suits the Federation of Brazil.

-*Year FE*: since our dataset (Huberts et al., 2025) offers several years, we will exploit time variation too. The methodological solution to make use of panel data is again FE. In fact, thanks to this instrument, we are able to isolate the variation among years and discard the magnitude of variation in a signle year.

This strategy isolates the effect of teachers quality on students test scores, which is commonly refered to as *outcome variable*.

Controls: The vector of controls is X_{smt}, while γ is the vector of coefficients. It represents the effects of control variables on our outcome variable. The selected controls are:

3.2 Intensive Margin

The second regression presented in the paragraph looks at the intensive margin. Therefore, it investigates the relationship between teachers quality and rates of failure. The equation used is the following:

$$Y_{smt}^{failure} = \beta_0 + \beta_1 TQ_{smt} + \beta_2 INFRA_{smt} + \gamma' X_{smt} + \mu_m + \lambda_t + \varepsilon_{smt}$$

[2]Cunningham (2021, chap. 2, p. 462) says: "I have a bumper sticker on my car that says "I love Federalism (for the natural experiments)". [...] United States is a never-ending laboratory. Because of state federalism, each US state has been given considerable discretion to govern itself with policies and reforms. Yet, because it is a union of states, US researchers have access to many data sets that have been harmonized across states, making it even more useful for causal inference."

Teachers Quality is measured as number of students per class *or* as teachers education. Again, identification relies on municipal FE, state FE and year FE.

4 Limitations

From the very first page, in this paper I adopted a transparet approach. Therefore, this section highlights all limitation of the analysis conducted so far. In this way, constructive criticism become active part of the scientific process.

I noticed three weakeness of the identification strategy:

- Omitted Variable Bias.
- Provide tests for observable implications in as many as possible narrow, focused, controlled circumstances (**<empty citation>**).
- Do not include control variables that are a consequence of the key IV(**<empty citation>**).

Omitted Variable Bias is a common burden of all empirical scientists. There is no way to include all possible sources of variation in the identification strategy and hampers the possibility of identifiyng the "perfect" causal mechanism. However, there are painkiller to this issue. They do not come from causal inference or econometrics. Instead, they come from institutional and qualitative knowledge of the topic under study.

Thus, we can reassure the reader of the reliability of the empirical result deepening the qualitative knowledge of the problem at stake.

A second important limit of this study comes from the fact that the more we generilise the results, the less we can be sure of what we state. Empirical tests have increadible internal validity. Nonetheless, they all suffer of a noticeable restriction, namely external validity. The identification strategy is solid only when we test the causal mechanism in data it was though to work for. When we appkly the identification strategy to external data, coming from a similar data generating process, the rusults tend to become fuzzy, and the ground of the analysis becomes slippery.

Eventually, we have another source of doubts about the research proposal. Let's imagin a state investing more in school with better tests scores or with lower failure rates. In this way, school managers or regional governments are forced to act in accordance with an incentives scheme.

However, if this were the case for Brazil, there would be a huge bais in the identification strategy used so far. This bias is due to the fact that the outcome variable (test score for example) shapes investments in the school infrastracture, that at the same time influences students performances.

Further exploration of the institutioinal setting can shed light on this issue.

Conclusion

Acknowledgements

Artificial intelligence-based tools were employed solely to improve linguistic clarity and grammar. No AI system contributed to the development of the research questions, theoretical framework or conclusions presented in this paper.

Bibliography

Books

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.

Southey, R. (2012). *History of Brazil*. Cambridge University Press.

Articles

Rubiane Daniele Cardoso de Almeida, Benjamin M. Tabak, & Tito Belchior Silva Moreira. (2023). Demographic aspects and regional income convergence in Brazil: A panel data approach. *CEPAL Review*, *2023*(139), 197–210. https://doi.org/10.18356/16840348-2023-139-10

Turmena, L., & Bitencourt, J. C. (2022). A reforma de 1ř e 2ř graus de 1971 e a reforma do ensino médio de 2017: Algumas aproximações. *Educ. Escr.*, *13*(1), e43895. https://doi.org/10.15448/2179-8435.2022.1.43895