A new dataset of average years of schooling in Brazil, 1925-2015

Júlia R. Walter[♣] Thomas H. Kang[♠]

17 March 2023

Abstract

This paper presents a new dataset of average years of schooling (AYS) for Brazil covering the years between 1925 and 2015, using demographic censuses and household surveys as benchmarks. The dataset is also broken down by gender (1925-2015), race/color (1925-2015), and states (1950-2015). In contrast to the Barro-Lee dataset, which reports an implausible AYS decrease in the 1970s, our dataset shows a gradual AYS increase in Brazil over time. We also show that women's AYS surpassed men's from 1983 onwards and that levels of inequality by race/color and state/region have remained relatively stable over time.

Keywords: average years of schooling, educational attainment, human capital, economic history of education

Resumo

Este artigo apresenta uma nova base de dados de anos médios de escolaridade (AME) entre 1925 e 2015, usando censos demográficos e pesquisas domiciliares como referência. Apresentam-se os dados por sexo (1925-2015), raça/cor (1925-2015) e estados (1950-2015). Em contraste com a conhecida série Barro-Lee, que relata uma diminuição implausível do AME na década de 1970, nossos dados mostram um aumento gradual dos AME no Brasil ao longo do tempo. Ademais, o indicador de AME feminina superou a masculina a partir de 1983 e que os níveis de desigualdade por raça/cor e estado/região mantiveram-se relativamente estáveis ao longo do tempo.

Keywords: anos médios de estudo, nível de escolaridade, capital humano, história econômica da educação

JEL Code: N36, H52, I25

Statements and Declarations:

The authors do not report any competing interests.

Author contributions:
Julia Walter: methodology, data collection, formal analysis and investigation, writing – original draft preparation;
Thomas Kang: conceptualization, methodology, writing – review and editing

♣ Software Engineering Associate, Quality Assurance - Accenture, Brazil. Contact: ju.rheiwalter@gmail.com
♠ Assistant Professor, Department of Economics and International Relations, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil. ORCID ID: 0000-0002-1964-3503. **Corresponding author: kang.thomas@gmail.com**

1 Introduction

Human capital is a widely recognized concept in the literature on labor markets and long-term growth since the early 1960s.[1] However, measuring human capital presents significant challenges. Although the notion encompasses various attributes, such as health and on-the-job training, formal schooling has received the most attention. Consequently, early studies on long run growth empirics used enrollment and literacy data as proxies for human capital (Mankiw, Romer and Weil 1992). However, the variable "average years of schooling of the working-age population" (hereafter AYS) became the most used in long run growth analyses shortly after (Barro 1999; Easterly and Levine 1998; Hall and Jones 1999; Rajan and Zingales 1998; Ramey and Ramey 1995; Sachs and Warner 1995). Barro and Lee (1993) facilitated this by creating an AYS worldwide dataset that initially covered data from 129 countries between 1960 and 1985 using five-year intervals.[2]

An important aspect of Barro and Lee (1993)'s contribution was methodological: they extensively used censuses and surveys (UNESCO, U.N. *Demographic Yearbooks*, and other sources) as benchmarks. After setting the benchmarks, the remaining gaps were filled with enrollment information using the perpetual inventory method (PIM).[3] In subsequent studies, Barro and Lee (2001, 1996, 2013) deployed an educational data projection methodology based on benchmarks. The latest methodological update is found in Barro and Lee (2013), and the most recent version of the dataset was in June 2018. The latest version encompasses data from 146 countries with estimates given at five-year intervals between 1950 and 2010 (this version will be called "BL dataset" hereafter).

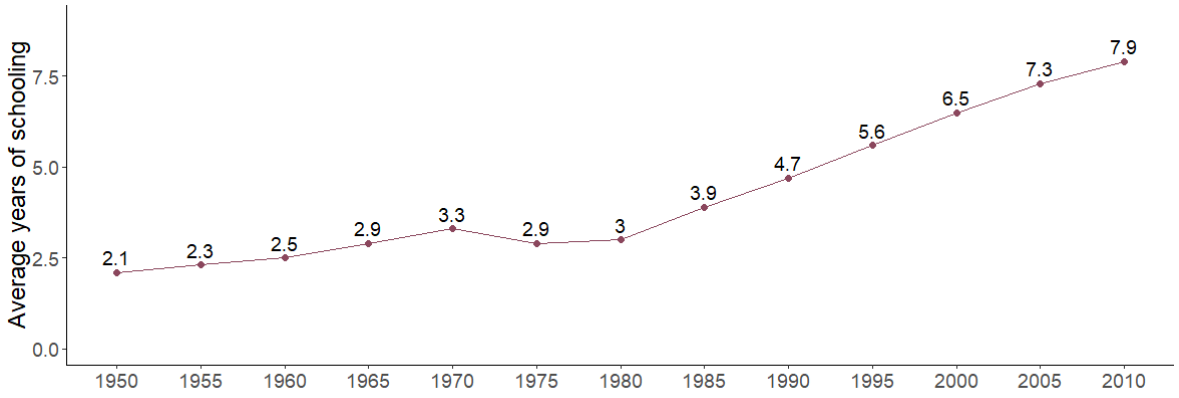
Despite the BL dataset's critical role in the literature, it still reports significant inaccuracies in individual country data (Speringer et al. 2015). This problem is particularly acute for Brazil – the largest Latin American country in terms of GDP, area, and population.[4] Brazil is an important case study of historical backwardness on the evolution of mass education: half of the adult population was illiterate in the mid-twentieth century according to official data. The Brazilian economy was strongly linked to slavery in the past, which may explain the persistent high levels of poverty and income inequality – issues that are

[1] Despite the early contribution of Mincer (1958), the notion of human capital became more widespread after Schultz (1961) and Becker (1964). Extensions of the work of Solow (1956) and other theoretical models incorporated the effects of human capital on long run growth (Aghion and Howitt 1992; Lucas 1988; Romer 1986, 1990).
[2] AYS also became an important variable in the human development literature, as the Human Development Index (HDI) composition demonstrates since at least 2010 (UNDP 2020).
[3] PIM is the method of transforming flow variables (in this case, enrollment) into stock variables (years of study) considering an educational gap.
[4] Barro and Lee (2013) not only presents the average years of schooling of the working-age population but also provides estimates of the population by level of education (graduates and non-graduates). One of the most generic problems with Barro and Lee's estimates is the reliability and accuracy of the data, in addition to the failure to decompose the levels of education between graduates and non-graduates (Speringer et al. 2015).

still associated to race/color and regional aspects. Problems in AYS estimates may lead to a distorted picture about the historical evolution of human capital and inequality in Brazil.

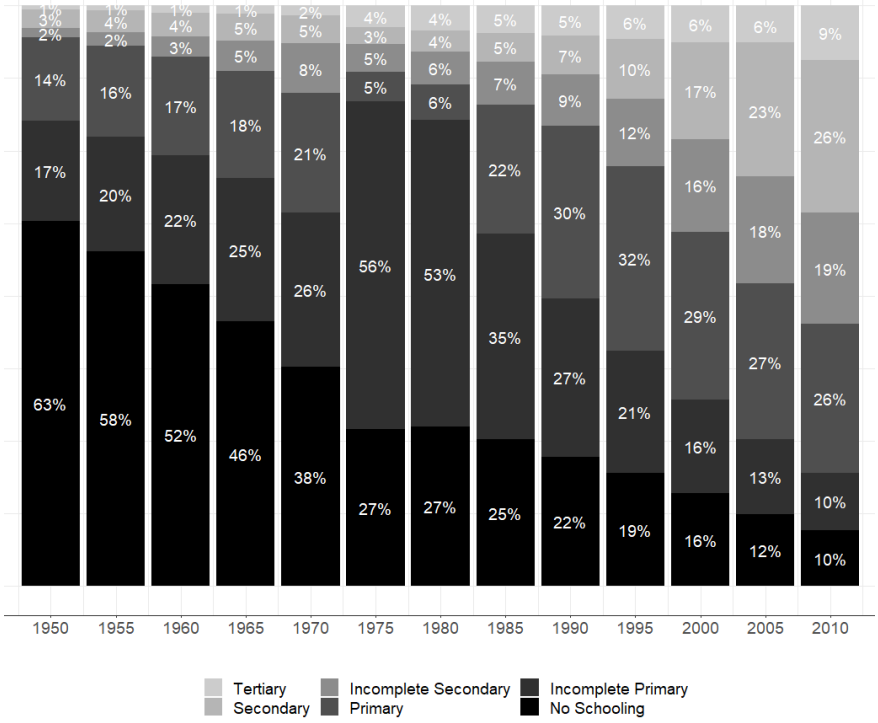
The Brazilian case also raise concerns about the BL methodology. According to the BL dataset, AYS decreased from 3.3 to 2.9 in Brazil between 1970 and 1975: a 12 percent drop in the stock of formal schooling (see Figure 1). The country would only have reached again its 1970 level in the early 1980s. Similar AYS decreases in other countries are usually associated with war periods (such as South Korea in the early 1950s). This was not the case for Brazil, a country that was experiencing its "economic miracle" between 1968 and 1973, with an average annual rate of economic growth of 11.1 percent.

Figure 1: Average years of schooling in Brazil, population aged 15 or over, 1950-2010, Barro and Lee (2018)



Source: Own elaboration, data from Barro and Lee (2018).

Figure 2: Distribution by education level according to Barro and Lee (2018), population aged 15 or over, Brazil, 1950-2010



Source: Speringer et al. (2015) and Barro and Lee (2018).

What could explain the unexpected decrease in AYS during the 1970s? Figure 2 provides estimates of population shares by schooling level (incomplete and complete) from the BL dataset. First, Brazilian data reports an abrupt rise in the population share that did not complete the primary level ("incomplete primary") during the 1970s. Secondly, it also shows a simultaneous decrease in the population share with "complete primary" during the same period. These changes are behind the sudden drop in AYS estimates. However, they are implausible: nothing indicates an increase in dropouts or retention in primary education during that period – in fact, retention indicators had been decreasing in the previous years (Kang, Paese and Felix 2021). These clear inconsistencies raise questions over the reliability of the BL estimates for the entire period, at least in the case of Brazil.

Bearing in mind these inconsistencies, this paper aims to present a new annual dataset of average years of schooling of the population aged 15 to 64 years in Brazil from 1925 to 2015, with discrimination by gender race/color, and regional subdivisions (states and regions from 1950 to 2015). To build the dataset, we used demographic censuses and household surveys (PNAD and PNAD-C) from the Brazilian Institute of Geography and Statistics (IBGE), in addition to the historical series of enrollment data from Kang, Paese and Felix (2021).

This study makes three contributions to the literature on the evolution of schooling in Brazil. First, we offer more accurate national-level data. Secondly, we provide an annual historical series rather than five-year or decennial estimates, by adapting the methodology of Földvári and Van Leeuwen (2009). Finally, our new dataset disaggregates data by gender, race/color, and regional sub-units. Moreover, our contribution may be useful for a wider audience for at least two reasons: (i) our methodology (or a similar one) may be applied to other countries, and (ii) case studies such as this one may call attention to inaccuracies in other countries' estimates.[5] A common cross-country methodology does not make comparisons more accurate if there are significant inconsistencies with national data.

Contrary to Barro and Lee's estimates, we show that AYS evolved with few shocks over the period in Brazil. Additionally, we demonstrate that Brazilian women surpassed men in education levels after 1983. Finally, race/color and regional inequalities in AYS showed a relatively stable behavior throughout the period. The Southeast and South regions consistently reported higher figures than the North region, whereas there was a significant AYS increase in Central-West and Northeast regions between 1950 and 2015.

This paper may also contribute to assessing AYS worldwide datasets. There are at least four other relevant international AYS datasets that use different estimation methods: Morrisson and Murtin (2009), Van Leeuwen and Van Leeuwen-Li (2014), Cohen and Leker (2014) and Lutz et al. (2018). However, significant divergences between these estimates raise doubts about which is the most appropriate (see Figure 4). Based on our estimates, we show that Lutz et al. (2018) reports the closest results (at least until 2010).

This paper is divided into five sections. Following this introduction, section 2 presents our methodology to estimate AYS in Brazil. Section 3 presents our results. Section 4 discusses the findings in detail, broken down in different categories, and compares them with those of previous studies. Section 5 concludes. A more comprehensive literature review is available in the working paper version of this article.

2 Materials and methods

Our estimates of AYS and educational attainment distribution were obtained using benchmarks as basic parameters, with the main sources being census and household surveys microdata. However, since census microdata is only available from the 1960 Census onwards, we used an alternative method for the period before 1960. Drawing on Lutz et al. (2007), we projected the 1960 census population backward in order to estimate AYS from 1925 to 1950. In order to produce an annual historical series of data between

[5] Our research is not designed for cross-country comparisons, because we do not have a complete database with the same methodology applied to the other countries.

censuses, we adapted a methodology used by Földvári and Van Leeuwen (2009) to ensure consistency between the shares of the population with different educational attainment levels (population without education, primary, secondary, and tertiary levels) and our AYS estimates.

2.1 Data and sources

As previously stated, our estimates of AYS for the working-age population between 1960 and 2000 were based on microdata from demographic censuses.[6] For the 2012-2015 period, we used a matching methodology with microdata from the PNAD-C. To estimate inter-census years from 1950 to 2012, we adapted the backward and forward estimation methodologies of Földvári and Van Leeuwen (2009). This involved using enrollment and population data that were interpolated by a cubic spline function.[7] Additionally, the methodology required information on the duration of each schooling level, AYS by schooling level, and population distribution by schooling level from the censuses between 1960 and 2000. By combining this information, we were able to estimate a dataset on attainment distribution and AYS.

We projected our 1960 data backward as a reference for years before 1960, as census microdata before that year is not digitally available. This methodology tracked cohorts over time and used population and enrollment data to estimate school attainment. This approach is useful for countries with limited data, such as Brazil, as it does not require a long series of enrollments like the estimation via enrollment flow/PIM. Furthermore, PIM does not ensure a school attainment distribution by level that is compatible with the AYS estimation (Morrisson and Murtin, 2009).[8] The projection methodology we employed only requires enrollment information for the school-age population, which is feasible in practical terms.

2.2 Benchmarks: demographic censuses and household surveys, 1960-2015

In order to use microdata, we need to convert the last grade each individual attended into years of schooling. However, censuses do not provide standardized information to carry out this conversion. Therefore, we have defined criteria for compatibility between censuses, which are detailed in Appendix B along with the sources (IBGE Census 1960-2000 and PNAD-C 2012-2015).[9]

[6] To carry out the estimation, it was necessary to reconcile the classification of the average years between the censuses as they differ in some details.
[7] R software functions used for interpolation belong to the *splinefun* package. See more in Kang, Paese and Felix (2021).
[8] For instance, to estimate the educational levels of the population aged between 15 and 64 via PIM in 1925, we would have to start with enrollment data from as early as 1868.
[9] IBGE census microdata are available in the repository of Instituto Base dos Dados (2020) with the exception of the 1960 census. In addition, Data Zoom provides support material with information for this compatibility, in addition to making compatible data via Stata. The 1960 census data comes from the universe questionnaire; 1970, 1980, 1991 and 2000 census data are from the sample. See "Instruções ao recenseador" in https://www.ibge.gov.br/.

There are various methods to convert attainment distribution by level into AYS. Barro and Lee (2013) estimate AYS by age group as a weighted average of the duration of each schooling level. Equation 1 represents AYS by age group in period t:

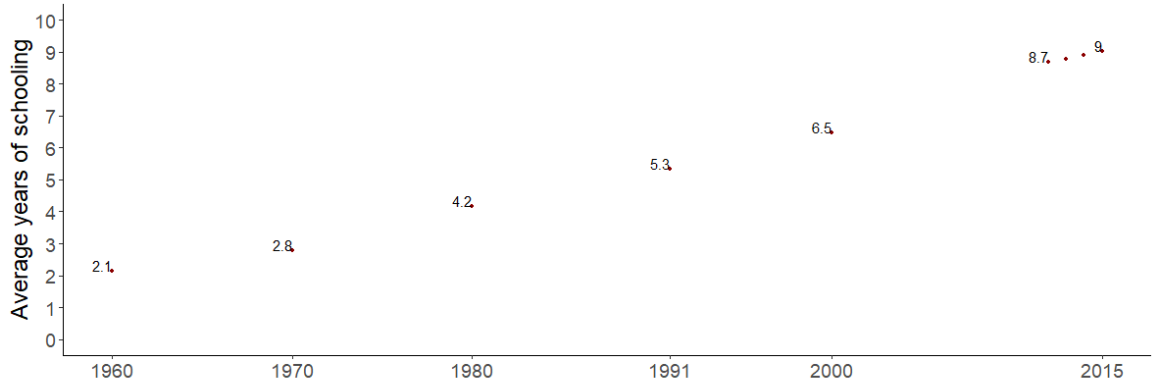
$$s_t^a = \sum_j h_{j,t}^a Dur_{j,t}^a \quad (1)$$

where Dur represents the duration of each schooling level; j corresponds to the schooling level; h is the age group share and its respective level of education, and a is the age cohort.[10]

In this study, we determined attainment by level (no schooling, primary, secondary, tertiary) in each benchmark. The distribution was calculated based on the years of schooling of each individual, enabling us to compute an accurate estimate of attainment by level and AYS. In contrast, Barro and Lee (2013) simply considered four years for "complete tertiary" and two years for "incomplete tertiary". However, setting the incomplete level duration as half of the complete level is questionable, especially in countries like Brazil, where historically high repetition rates have been observed (Ribeiro 1991). We acknowledge that microdata is not available in several countries, which may explain such choices in international studies.

Figure 7 displays the benchmarks: AYS from censuses (1960–2000) and PNAD-C (2012 and 2015). Figure 8 illustrates the distribution by schooling level (complete and incomplete) in each census. Individuals who did not report their last completed grade were excluded from the database.

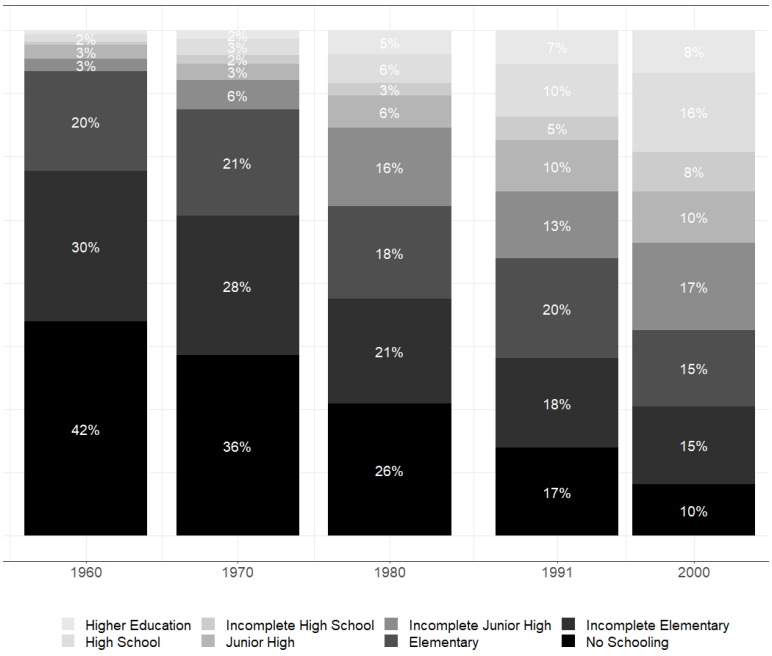
Figure 7: Benchmark estimates of average years of schooling, population aged between 15 and 64 years, Brazil, 1960-2015



[10] For example, the average years of schooling of the age cohort from 15 to 19 years (a = 15-19 years) represents the sum of the educational distribution of this age group multiplied by the respective duration of the level of education. By hypothesis, assuming that in 1960: 53% of the cohort had incomplete EF; 3% complete EF; 3% MS incomplete; 1% MS complete; 0.5% ES incomplete; 0.2% ES complete. Furthermore, considering the duration of each level of education: 4 years; 8 years; 10 years; 11 years; 12 years; 14 years. From these data, applying Equation (1), an average of 2,858 years of schooling is obtained for the cohort aged 15 to 19 years. Likewise, the average schooling of the population aged 15 to 64 years can be found from the weighted average of this result applied to each age cohort.

Source: Own preparation based on microdata from the IBGE Censuses (1960 – 2000) and data from PNAD-C (2012 – 2015).

Figure 8: Educational attainment distribution according to IBGE microdata, population aged between 15 and 64, Brazil, 1960-2000[11]



Source: Own elaboration, microdata from the IBGE Censuses (1960 – 2000).

2.3 Estimation: 1925-1950

Given the scarcity of data in this period, we monitored cohorts from 1925 to 1950 through a methodology adapted from Lutz et al. (2007). We made from the following assumptions:

a) Immigration does not affect educational patterns: schooling differences between immigrants and locals are not considered;

b) We assume different mortality patterns by age group but homogeneous mortality between groups by schooling level. In other words, schooling level and mortality rate are independent variables. This assumption seems acceptable for earlier periods (Cohen and Soto 2007);

c) Educational attainment remains constant for the population aged 25 years or more;

[11] Incomplete primary = population with 1 to 3 years of schooling; complete primary = 4 years of schooling; incomplete secondary = 5 to 7 years of schooling; complete secondary = 8 years of study; incomplete secondary = 9 to 10 years of education; complete secondary = 11 years of study; complete and incomplete tertiary = 17 years of education.

d) Duration by schooling level from 1925 to 1960 is equivalent to the estimated durations in 1960 (Dur). In other words, we assume that the proportion of people with "incomplete level" is the same during this period;

e) We assume a constant gross enrollment ratio (GER) from 1910 to 1933 for secondary and tertiary education. For primary education, we used the number of students enrolled from grades 1 to 4 to calculate the gross ratio (1910-1933).[12]

The methodology adopted is similar to those used by KC et al. (2010) and Lutz et al. (2007). Initially, the population aged 25 and over from the 1960 census is taken to project the cohorts into the past. For example, the proportion of the 40-year-old cohort that had completed primary education in the 1960 census will be identical to the proportion of the 30-year-old cohort that had completed the primary level in 1950. This methodology was applied to all non-school cohorts, enabling estimation back to 1925. Assuming that changes in school attainment only happen among the school-age population, the educational attainment of the population aged between 15 and 25 years was measured through the GER, taking a lag into account. For example, the cohort aged 20 to 24 years in 1950 attended primary education in 1935 (15 years lag), attended the secondary level in 1940 (10 years lag), and achieved the age group to attend the tertiary level in 1950 (no lag). In this case, lags are considered for listing which GER for the period is equivalent to the educational distribution of this cohort. The same procedure is applied to the age group from 15 to 19 years old who in 1950 were attending secondary education and who attended primary education in 1945 (10 years lag).

To estimate the 1950 population, we utilized enrollment data from Kang, Paese, and Felix (2021) for the school-age cohorts (15 to 24 years) and data from the 1960 census for the population aged 25 to 64 years. We determined the final distribution by applying a weighted average to the educational distribution results of the population aged 15-19, 20-24, and 25-64 years. To convert these data into AYS, we multiplied the educational distribution (primary, secondary, tertiary) by the respective duration of the level of education in 1960, including non-graduates, which is a constant variable.

In addition to the methodological advantages mentioned earlier, this technique does not rely on grade repetition data, which is convenient given Brazil's lack of reliable historical data on repetitions and dropouts over an extended period (Ribeiro 1991).[13] Furthermore, due to the low availability of data on net enrollment, GER was used to estimate the educational distribution of the school-age population. This choice

[12] The reference year considered as a constant value is the year 1933, as it is the first data observed in the historical series of enrollment of Kang, Paese and Felix (2021) for levels of education.

[13] Benchmark estimation allows for greater precision in the estimates. The methodology of tracking cohorts over time (1925 to 1950) avoids the use of a very long enrollment series. The PIM-based measurement technique of Földvári and Van Leeuwen (2009) makes it possible to create an annual series of AYS.

tends to overestimate the data, as does the use of duration by the level of education from 1960 as a constant parameter for converting educational distribution into AYS. On the other hand, as mortality rates change over time, there is a tendency to underestimate the variables, since the population aged 25 and over in 1960 is used as a reference for estimating the population aged 25 to 64 years from 1925 to 1950.[14]

The perpetual inventory method (PIM) was also tested, but it only proved effective for more recent time series intervals.[15] This methodology is challenging to replicate for earlier periods, as it requires an extensive historical series of enrollments. Furthermore, we only have scattered information on enrollment broken down into subgroups (such as by gender or state).

2.4 Inter-Census Periods, 1950 - 2012

Between 1950 and 2012, we applied the adapted estimation methodology based on the perpetual inventory method (PIM) of Földvári and Van Leeuwen (2009), which is a modified version of the Barro and Lee (1993, 2001) approach, to generate an annual series of AYS. The Barro and Lee method does not account for differential mortality and dropouts, which may introduce bias in estimates in more recent periods (Földvári and Van Leeuwen 2009). Neglecting mortality differentials tends to underestimate the AYS of the older population, while not considering dropouts can overestimate it. To address this, we assumed that the percentage of dropout and differential mortality is constant, and removed the bias by taking the average of forward and backward estimations of equidistant benchmarks. [16]

Equation 2 presents the forward estimation methodology:

$$h_{0,t} = H_{0,t}/L_t = h_{0,t-i}[1 - (L15_t \cdot i/5 \cdot L_t)] + (L15_t \cdot i/5 \cdot L_t).(1 - PRI_{t-i})$$

$$h_{1,t} = H_{1,t}/L_t = h_{1,t-i}[1 - (L15_t \cdot i/5 \cdot L_t)] + (L15_t \cdot i/5 \cdot L_t).(PRI_{t-i} - SEC_t)$$

$$h_{2,t} = H_{2,t}/L_t = h_{2,t-i}[1 - (L15_t \cdot i/5 \cdot L_t)] + (L15_t \cdot i/5 \cdot L_t).SEC_t - (L20_t \cdot i/5 \cdot L_t).HIGH_t$$

$$h_{3,t} = H_{3,t}/L_t = h_{3,t-i}[1 - (L15_t \cdot i/5 \cdot L_t)] + (L15_t \cdot i/5 \cdot L_t).HIGH_t \quad (2)$$

Equation 3, in turn, shows the backward method:

$$h_{0,t-i} = \left(h_{0,t} - (L15_t \cdot i/5 \cdot L_t).(1 - PRI_{t-i})\right)/[1 - (L15_t \cdot i/5 \cdot L_t)]$$

$$h_{1,t-i} = \left(h_{1,t} - (L15_t \cdot i/5 \cdot L_t).(PRI_{t-i} - SEC_t)\right)/[1 - (L15_t \cdot i/5 \cdot L_t)]$$

$$h_{2,t-i} = \left(h_{2,t} - (L15_t \cdot i/5 \cdot L_t).SEC_t + (L20_t \cdot i/5 \cdot L_t).HIGH_t\right)/[1 - (L15_t \cdot i/5 \cdot L_t)]$$

[14] However, the population's life expectancy possibly tends to be lower from 1925 to 1950.
[15] Similar to the methodologies adopted by Lee and Lee (2016), Nehru, Swanson & Dubey (1995), Földvári and Van Leeuwen (2013). In section 3.4 we use a modified PIM technique for inter-census periods.
[16] See Földvári and Van Leeuwen (2009). The sum of the educational distributions of our estimates did not achieve 100%. The maximum difference found in the sum of the educational distribution of the total population was 1.16 percent (the average difference being 0.33 percent). Because of this, there was an adjustment in the educational distribution: the difference was distributed equally between the levels of education.

$$h_{3,t-i} = \left(h_{3,t} - (L20_t \cdot i/5 \cdot L_t).HIGH_t\right)/[1 - (L15_t \cdot i/5 \cdot L_t)]$$

(3)

For both sets of equations, h represents attainment per educational level (with indices 0, 1, 2, and 3 representing no education, primary, secondary, and tertiary levels, respectively). H corresponds to the population by level of education; i is the number of years between the year to be estimated and the benchmark; L is the population aged between 15 and 64 years; $L15$ is the population aged between 15 and 19; $L20$ is the population aged between 20 and 24. PRI, SEC, and $HIGH$ represent the GER in primary, secondary, and tertiary education, respectively.

To estimate AYS in 1965, we applied the forward methodology for 1960 and backward for 1970, and then took the average of the results. Thus, we obtained the percentage of the population in 1965 who had no schooling, attended primary, secondary or tertiary education during the period. From the 1965 educational distribution, it was possible to calculate the 1963 educational distribution and so on.

A methodological adaptation was needed to determine the duration of each schooling level for the conversion of the attainment information into AYS. Földvári and Van Leeuwen (2009)'s methodology cannot distinguish between the population who have completed their level of education and those who did not. To compensate for this lack of information, the duration of each educational level was estimated based on the duration reported by the nearest censuses. For insntace, to calculate the average years of schooling in 1965, each educational distribution was multiplied by an educational level duration corresponding to the weighted average duration of the 1960 and 1970 censuses. If the duration of the educational level of the primary level in 1970 was 6 years and in 1960 it was 5 years, the weighted average of the two durations was applied (5.5 years – with a weight of 0.5, considering that 1965 is equidistant to 1970 and 1960; if it were 1964, 1960 would have a larger weight).

Based on the adapted methodology, equidistant points were created between the benchmarks before and after the estimated year whenever possible, as this reduces biases between backward and forward estimations.[17] For instance, in estimating educational distribution in 1972, the census of 1970 and the distribution of 1974, which have already been estimated, can be used. In this case, 1972 is equidistant from 1970 and 1974. However, equidistant estimates were not always possible due to the distance from the benchmarks. In such cases, larger weight was assigned to the backward estimates, which results in

[17] Backward estimation tends to overestimate, while forward tends to underestimate.

overestimation (e.g., to estimate 1952, 1950 and 1955 were used as benchmarks, a distance of two and three units, respectively). Nonetheless, this difference in distances never exceeded one.[18]

2.5 Estimation by subgroups

We estimated AYS by gender (men and women), race/color (yellow, white, colored, and black), and subnational divisions (states) using our methodology. For the gender case, we added additional steps. As enrollment rates by gender were not available for earlier years, we did not use them. Consequently, our dataset may slightly underestimate men's schooling and overestimate women's in earlier periods, as girls reportedly had lower enrollment rates in the early 20th century. Nevertheless, this is a minor issue since the gender gap was insignificant by the mid-century.[19]

Estimates by race were made similarly, except for 1970 as there is no information on race/color in this census. To address this gap, we applied the methodology of Földvári and Van Leeuwen (2009) between 1960 and 1980. Moreover, GER estimates by race/color in the period 1925-1960 were based on the 1960 census.[20] State-level estimates were based on the classification of states in 1940. Furthermore, we aggregated states to report estimates by regions (North, Northeast, Center-west, Southeast, and South).[21]

2.6 Population Estimates and Enrollment Data

We estimated the population by state for the period 1950-2015 for age ranges of 5-14, 15-19, 20-24, and 60-64 years, using a cubic spline function interpolation from demographic censuses and PNAD-C data. National estimates were obtained by summing the state-level estimates. For the population from 1925 to 1950, we used IBGE data and interpolated between years. The female population was based on microdata from 1960 onwards, while before 1960, we calculated the proportion of women over the total population,

[18] In some years when data were not equidistant, there was overestimation: 1952, 1962, 1972, 1957, 1967, 1977, 1982, 1985, 1986, 1995, 2001, 2004.

[19] Furthermore, in an analysis of data from Kang, Paese and Felix (2021) from 1933 to 1970 the proportion of female enrollments about the total was, on average, 49.1% in primary education, with the lowest percentage, around 48.2%, in 1933. From 1956 to 1970 the average was 49.3% for elementary education. These data indicate that the use of similar gross enrollment ratios for men and women is a valid hypothesis. On the relative gender equality in enrollment in Brazil and Latin America, see also Frankema (2008) and Kang (2010).

[20] According to the 1960 Census, the share of yellow, colored/black, and white children among students attending the old primary school were 1%, 32%, 67%, respectively. In secondary education, these shares were 2%, 8%, 90%; while in the tertiary level, these shares were 2 %, 4%, and 94%. These shares by racial groups were applied for the period 1925-1960 and then we generated gross attendance ratios. Considering that 88% of the enrolled population attended school in 1960, this proportion was used to convert gross attendance into GER (for all races). For secondary and tertiary education, we used gross attendance since a small proportion of people attended these schooling levels.
That is, from 1960 to 1925 the GER/gross attendance by race/color was estimated, but from 1960 to 2015 the same gross rate of the total population was used (regardless of race/color) because the model proved to be little sensitive to the variation in enrollment in this period, mainly due to the use of a benchmark.

[21] The exception was the territory of the current state of Tocantins, considered here as part of the Center-west region. We used states as defined in 1940, when the current Tocantins still belonged to the state of Goiás. The Federal District was not included in the aggregations to avoid distortions.

with equivalent proportions for all age groups.[22] The male population was obtained by residual. We followed a similar procedure to estimate population by race.

Enrollment statistics were obtained from Kang, Paese, and Felix (2021). To estimate GER by schooling level, we used the age ranges of 5-14 for primary education, 15-19 for secondary, and 20-24 for tertiary education.[23] There is an enrollment series for the old primary education (grades 1-4) from 1872 to 2013, with some missing periods filled by interpolation. Between 1900 and 1933, GER in the new primary level (grades 1-8) were based on data for the old primary education (grades 1-4), leading to an underestimation bias. As information for secondary and tertiary education is only available from 1933 onwards, we assumed a constant GER in these schooling levels from 1900 to 1933.

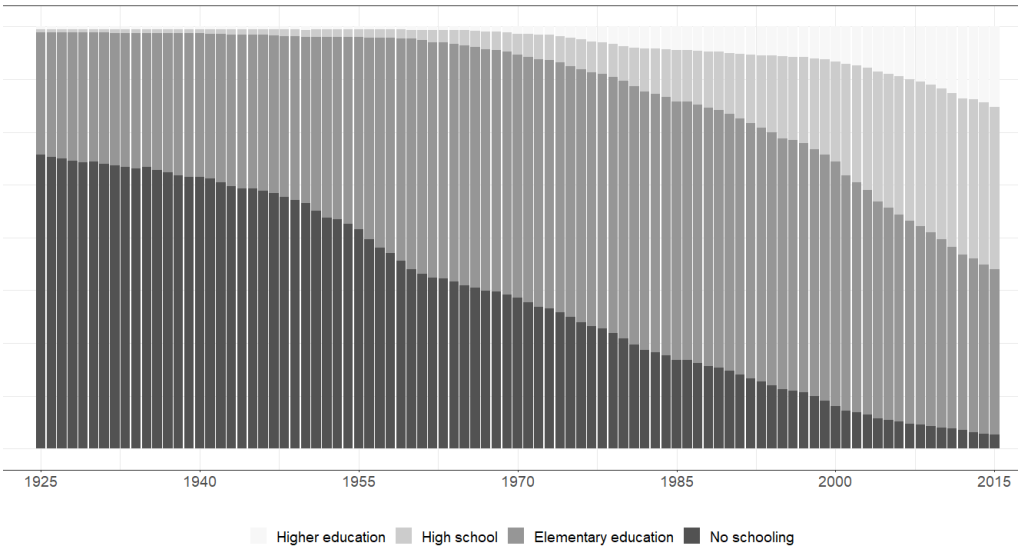
3 Results

3.1 Total population from 15 to 64 years old

Figure 9 presents our annual historical series of the attainment distribution for the Brazilian population aged between 15 and 64 years. The evolution of the attainment distribution appears consistent over time. The few episodes of more abrupt change are plausible, such as the increase in the proportion of people who completed primary education or had incomplete primary education in the mid-1950s, which probably resulted from the acceleration of the enrollment rate growth in primary education after 1945.[24] After applying Equation 1 to this attainment distribution data, we obtained our AYS estimates, presented in Figure 10. Figures 9 and 10 present the main results of this paper.

[22] The proportion of women over the total population per census year is available in the 1950 census report.
[23] Kang, Paese, and Felix (2021) calculated enrollment rates for primary education (EF) using the age group 7-14 and for secondary education (EM) using the age group 15-17. In our study, we opted to use the same age bands as Földvári and Van Leeuwen (2009) for consistency.
[24] From 1950 onwards, there is a change in methodology due to the low number of benchmarks. Therefore, this stabilization of the population without education could be a problem caused by the methodological change. However, if the illiteracy rates registered in 1920, 1940, and 1950 censuses of the population aged 15 years or more are analyzed (64.9%, 55.9%, 50.5%), the values do not differ much from the population without education estimated in these periods for the population aged 15 to 64 years (69% in 1925, 64% in 1940, 58% in 1950). This fact gives us greater confidence in our estimates.

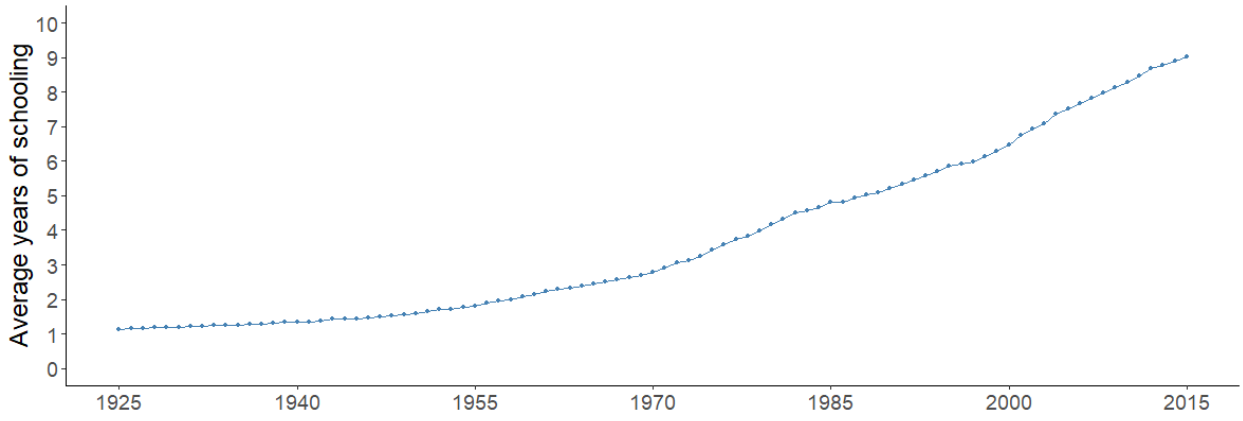
Figure 9: Educational attainment distribution, population aged between 15 and 64 years, Brazil, 1925-2015



Source: Own elaboration, see Section 3.

Note: The classification also includes those who did not complete the level. For example, those who did not complete primary education are also included in the primary education category.

Figure 10: Average years of schooling, population aged between 15 and 64 years, Brazil, 1925-2015.



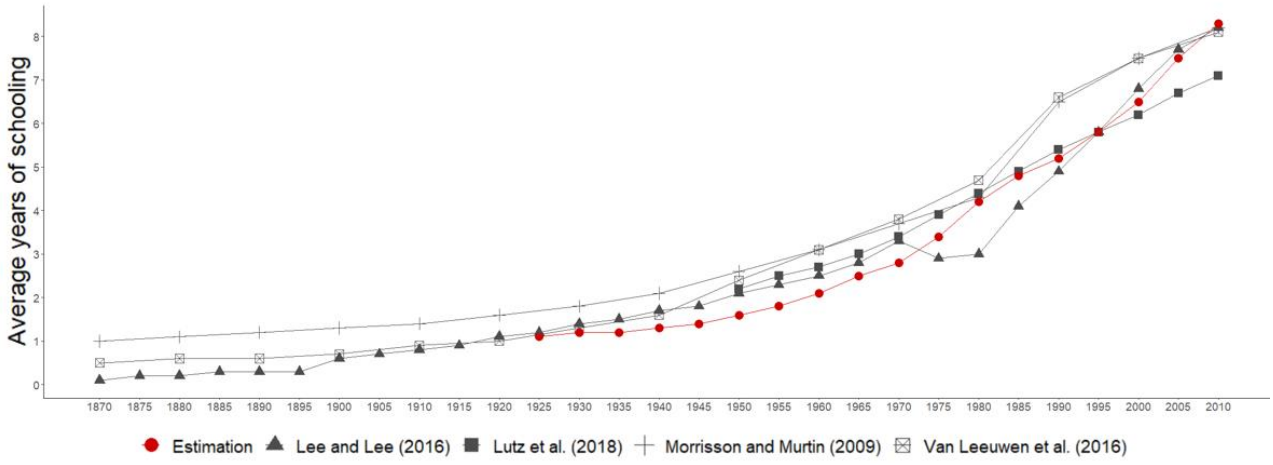
Source: Own elaboration, see Section 3.

4 Discussion

4.1 General discussion

Figure 11 reveals that our AYS estimates differ significantly from those of Lee and Lee (2016) and Morrisson and Murtin (2009), who report higher estimates. Even though Van Leeuwen and Van Leeuwen-Li (2014) and Lutz et al. (2018) used a different population range (those aged 15 years or older), they also provide higher estimates. If these studies adopted the population aged between 15 and 64 years, as in our study, they would probably report even higher results since older generations attained lower levels of schooling.

Figure 11: Average years of schooling according to different studies, Brazil, 1870-2015



Source: Own elaboration, based on data from Lee and Lee (2016), Morrisson and Murtin (2009), Van Leeuwen, Van Leeuwen-li and Földvári (2016) e Lutz et al. (2018). See section 3.

Table 2: Estimated AYS compared to other studies in relative terms (our series = 1.00), Brazil, 1920-2000

Year	Lee and Lee (2016)	Morrison and Murtin (2009)	Van Leuween et al. (2016)	Lutz et al. (2018)
1950	1.33	1.61	1.48	1.39
1960	1.18	1.43	1.44	1.27
1970	1.18	1.32	1.35	1.23
1980	0.73	1.03	1.12	1.06
1990	0.95	1.26	1.28	1.04
2000	1.05	1.16	1.17	0.97
2010	0.99	0.99	0.98	0.85
R-squared	0.94	0.96	0.96	0.98

Source: Own elaboration, see first line of Table 2.

Note: Our estimate is the reference for calculating R^2 and comparing studies in relative terms.

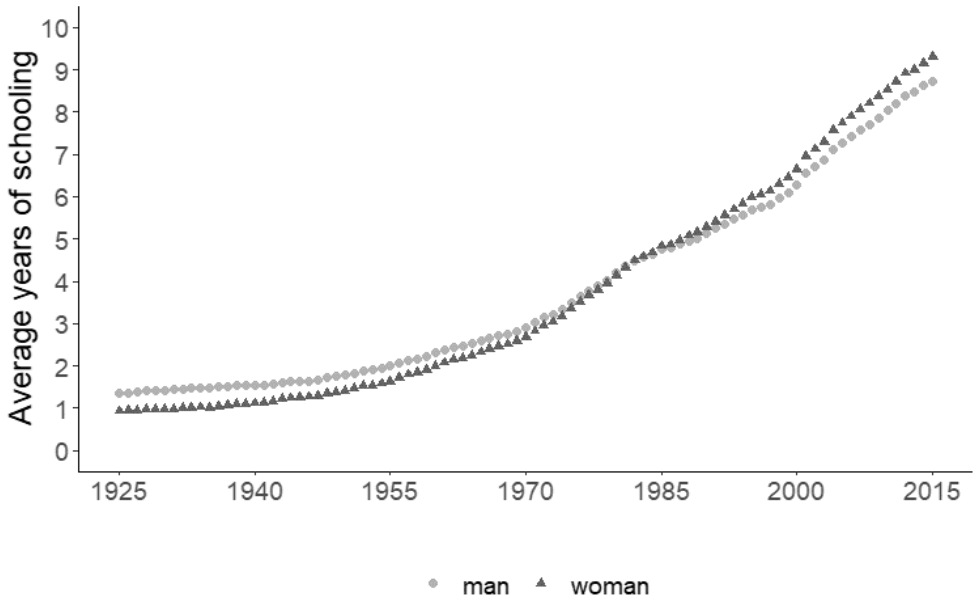
One of the justifications for this overestimation is that Lutz et al. (2018) and Lee and Lee (2016) use Equation 1 to convert educational distribution into AYS, but not with a duration of schooling that encompasses the population with incomplete education in the country. In the cases of Van Leeuwen and Van Leeuwen-Li (2014) and Morrisson and Murtin (2009), both tend to overestimate from 1960 onwards because they use Cohen and Soto (2007) as a reference (section 3 explains the issues of Cohen and Soto's estimates). Table 2 shows that Lutz et al. (2018) is the one that best fits our estimates, despite clear inaccuracies in the estimates after 2010. Our results raise questions about the reliability of some international datasets, which are prone to errors given the different organization of school systems across countries. In fact, using a common methodology for different countries may deliver estimates that are far from reliable – even for comparative purposes.

4.2 AYS: men and women

Figure 12 shows the AYS gap between men and women in Brazil. The 1991 Census had already revealed that women had surpassed men in terms of educational attainment (Melo and Thomé 2018). According to our annual AYS estimates, the turning point in favor of women was 1983.[25] A factor that may be associated with the increase in women's educational attainment is the declining fertility rate in Brazil, which fell from 4.35 children per woman in 1980 to 2.89 in 1991. According to national censuses from 1960 to 2000, the greatest fall in the fertility rate occurred in the 1980s. It is difficult to define the direction of causality: the drop in the number of children may have enabled women to enter the labor market and attain higher educational levels, or the opposite may have occurred.[26]

Figure 12: Average years of schooling by gender, population aged between 15 and 64 years, Brazil, 1925-2015

[25] This milestone is close to 1979, when the United Nations General Assembly declared that women enjoyed the same rights and duties as men (The United Nations, 1988).
[26] Fertility rate IBGE census data: 1950 = 6.21 children per woman; 1960 = 6.28 children per woman; 1970 = 5.76 children per woman; 1980 = 4.35 children per woman; 1991 = 2.85 children per woman; 2000 = 2.38 children per woman.



Source: Own elaboration, see Section 3.

Lee and Lee (2016) also estimated AYS by gender in Brazil and found that women's AYS surpassed men's at some point between 1985 and 1990. Lutz et al. (2018) found that Brazilian women's average schooling surpassed that of Brazilian men in 1980. The U.S. data can provide us with some perspective. According to Lutz et al. (2018), women already had more schooling than men in the US in 1950. Although women were overtaken by men in 1975, they regained the lead in the US in 2000..

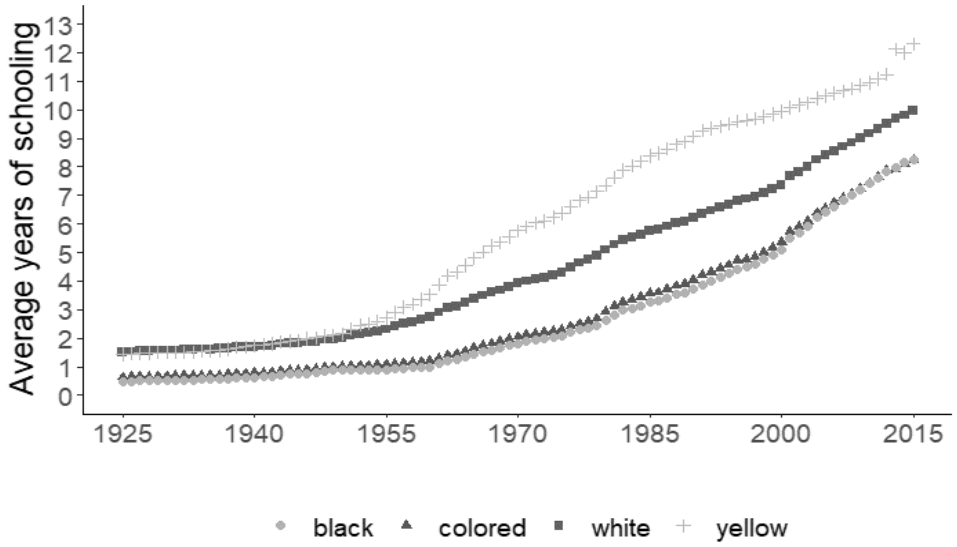
4.3 AYS: race/color

The evolution of AYS by race/color reveals that the gaps in education levels persist and are more stable in Brazil. The Asian Brazilian (*amarela*) population has attained a considerably higher education level than the other groups: approximately 50% of the population attended or completed tertiary education in 2015.[27] However, this group is a minority in the Brazilian population. However, this group represents a minority of the Brazilian population. The African Brazilian groups (*parda* or *preta*), which comprise the majority of Brazilians, reported lower AYS levels. From 1979 to 1992, our series shows the greatest absolute black-white gap in terms of AYS, but there has been a slow convergence between racial groups in the last few decades (see Figure 13).[28] This period coincides with the time when the country was facing

[27] Brazilian surveys and censuses still use words such as *amarelo* (yellow), *negro* (black), or *pardo* (colored), instead of Asian Brazilian or Afro Brazilian.
[28] In 1979, there was a difference of 2.2 fewer years of schooling for the colored population concerning the white population and a difference of 2.5 for the black population. However, the percentage variation between whites and coloreds/blacks has decreased over time, albeit slowly.

foreign debt and inflation crises (known as "the lost decade"). In 2015, the white, colored, black, and yellow groups had attained 9.9, 8.3, 8.2, and 12.3 years of schooling, respectively.

Figure 13: Average years of schooling by race/color, population aged between 15 and 64 years, Brazil, 1925-2015



Source: Own elaboration, see Section 3.

Note: The break c. 2013 in the AYS of the Asian Brazilian population is likely caused by a methodological change in the national household survey (from PNAD to PNAD-C). Brazilian surveys and censuses use words such as *amarelo* (yellow), *negro* (black), or *pardo* (colored), instead of Asian Brazilian or Afro Brazilian.

4.4 AYS: states and Regions

The historical series of AYS by regions (see Figure 14) reveals differences consistent with what we know about regional economic inequalities, with the Southeast, South, and Central-west having higher incomes than the North and Northeast.[29] he literature reports that the Northeast Region of the country has been relatively poorer than the southern regions since at least the mid-19th century (Baer 1964; Barros 2012; De Carvalho Filho and Monasterio 2012; Denslow 1973; Leff 1972; Monasterio 2010; Naritomi, Soares, and Assunção 2012; Pereira 2021; Reis 2014; Williamson 1965). In 1950, when our historical series by regions begins, the per capita income of the Northeast Region was close to a quarter of the income of the Southeast. Several studies have emphasized the role of human capital in regional inequalities in Brazil (Barros 2012; Oliveira and Silveira Neto 2016; Pessôa 2001). Kang, Paese, and Felix (2021) have shown that there were substantial differences between regions in terms of gross enrollment ratios (GER), which

[29] These regions (Southeast, South, Central-west, Northeast and North) are officially recognized by the IBGE.

only became more convergent towards the end of the twentieth century. Similarly, our series shows the persistence of inequalities in terms of average schooling across regions.

The evolution of the Central-west region is notable, as it surpassed the North region in terms of AYS in 1977. This rapid educational progress may be attributed, at least in part, to the creation of Brasília in 1962, although the Federal District is not included in this analysis to minimize distortions. In addition to the externalities of the new capital to surrounding areas, government incentives led to the occupation of the Central-west, which in turn led to the expansion of the agricultural frontier from the 1970s onwards. A large part of this migration came from the South and Southeast, the most educated regions at the time, which probably explains the expansion of schooling in the Central-west. Although the Northeast has not surpassed any other region in terms of AYS, the region had the highest coefficient of variation from 1950 to 2015.[30] This result may be associated with the increase in the number of children in schools since the 1980s, the expansion of school coverage, and a decrease in the illiteracy rate (Castro 1999).[31] Starting in 1994, the "Basic Education Project for the Northeast" (Projeto da Educação Básica para o Nordeste) was created, which led to investments of around US$ 800 million. Moreover, the Fund for Maintenance and Development of Primary Education (Fundef) increased education funding, particularly in poorer regions (Castro 1999).[32] However, the more developed regions, Southeast and South, have maintained the lead since the beginning of the series by regions and states in 1950.

Figure 15 provides a detailed view of regional inequalities. From 1950 to 1980, there was a significant dispersion between the state with the highest level of AYS, Rio de Janeiro (RJ), and the second-place state, São Paulo (SP).[33] However, RJ was surpassed by SP in 2002. An emblematic case is Rio Grande do Sul (RS), which was on par with SP from 1950 to 1980 but began to decline in 1982. In 2015, RS had reached 9.2 years (the 5th most educated state), while SP had already reached 10.0 years. With a few exceptions, the distribution of AYS between states remained largely unchanged throughout the series, as depicted in Figures 15 and 16. [34]

[30] Available data confirm that the greatest variation in schooling was observed between 1965 and 1985. Northeast (NE) and Central-west (CW) regions reported higher coefficients of variation, approximately 30%, and 29% respectively. South (S) and North (N) regions reported a coefficient of 22%, whilst the Southeast reported 18% in the period. From 1950 to 2015 the coefficient of variation of the S, SE, N, NE, and CW regions was respectively: 45%, 43%, 52%, 65%, 59%.

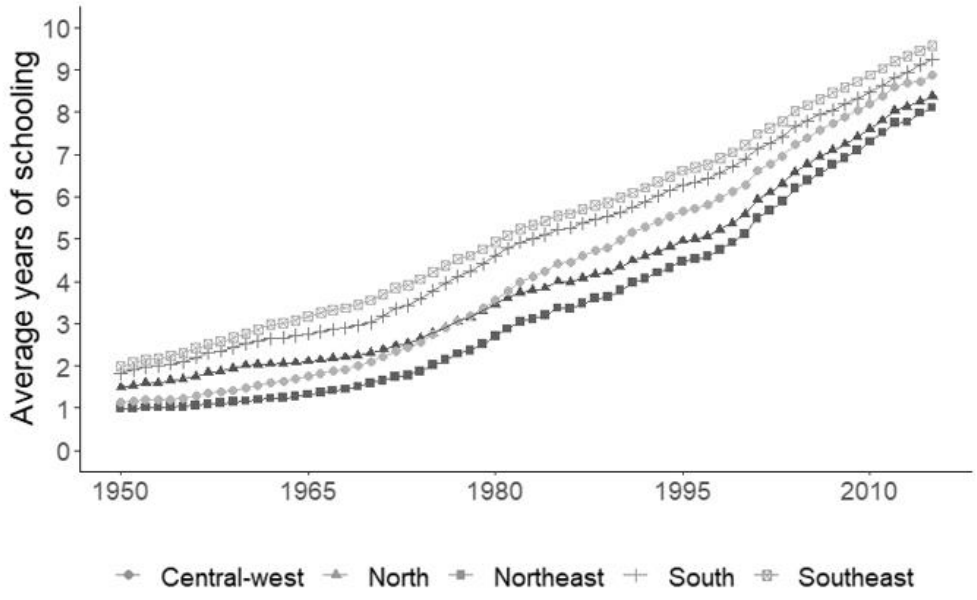
[31] In 1980, the Northeast had an illiteracy rate of 45.5% (population 15 years old or more), 2.7 times greater than the Southeast, this number was reduced to 28.7% in 1996, but with a difference of 3.3 times, compared to the Southeast, IBGE data (Castro 1999).

[32] According to Castro (1999), the Basic Education Project for the Northeast was supported by the World Bank. Its investments reached around R$ 800 million in nine states over 6 years. Fundef, on the other hand, allowed a significant increase in the average salary of teachers in municipal and state public schools, mainly in the North and Northeast regions.

[33] Years of schooling: 1950, RJ = 2.76 and RS = 2.05; 1960, RJ = 3.78 and SP = 2.89; 1970, RJ = 4.29 and SP 3.78; 1980, RJ = 5.73 and RS = 5.13; 1991, RJ = 6.81 and SP = 6.29.

[34] Except for GO, which increased its schooling pairing with Espírito Santo and Pará, which had a similar level of education to Mato Grosso in 1950 and which became similar to Paraíba in 2015 (decrease in education).

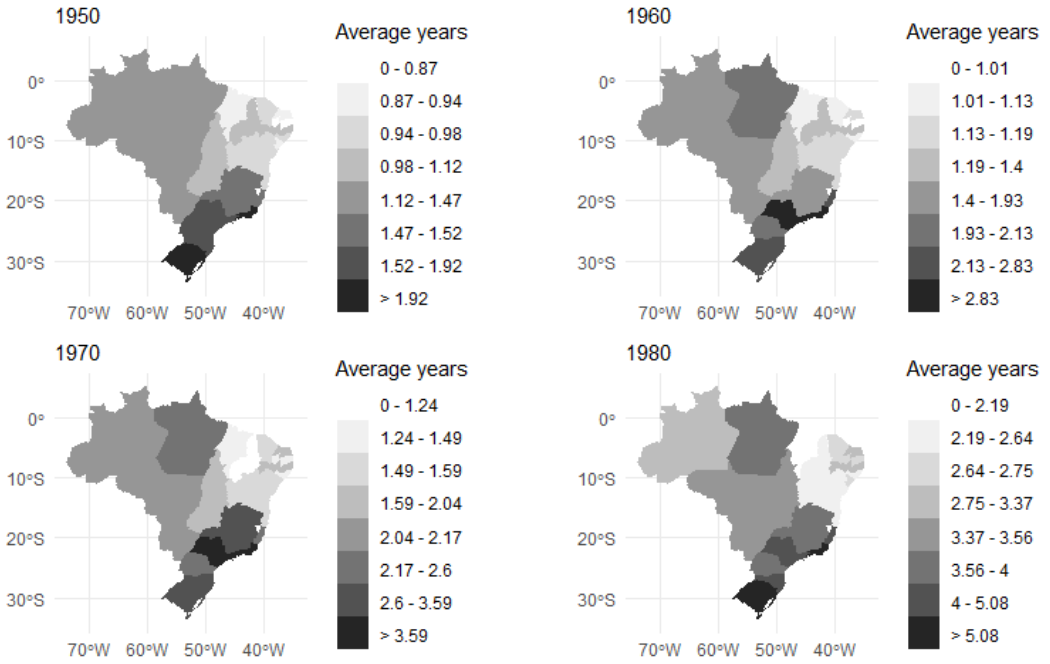
Figure 14: Average years of schooling by Brazilian region, population aged between 15 and 64 years, Brazil, 1950-2015



Source: Own elaboration, see Section 3.

Note: The Federal District was not counted and Tocantins belongs to the Midwest.

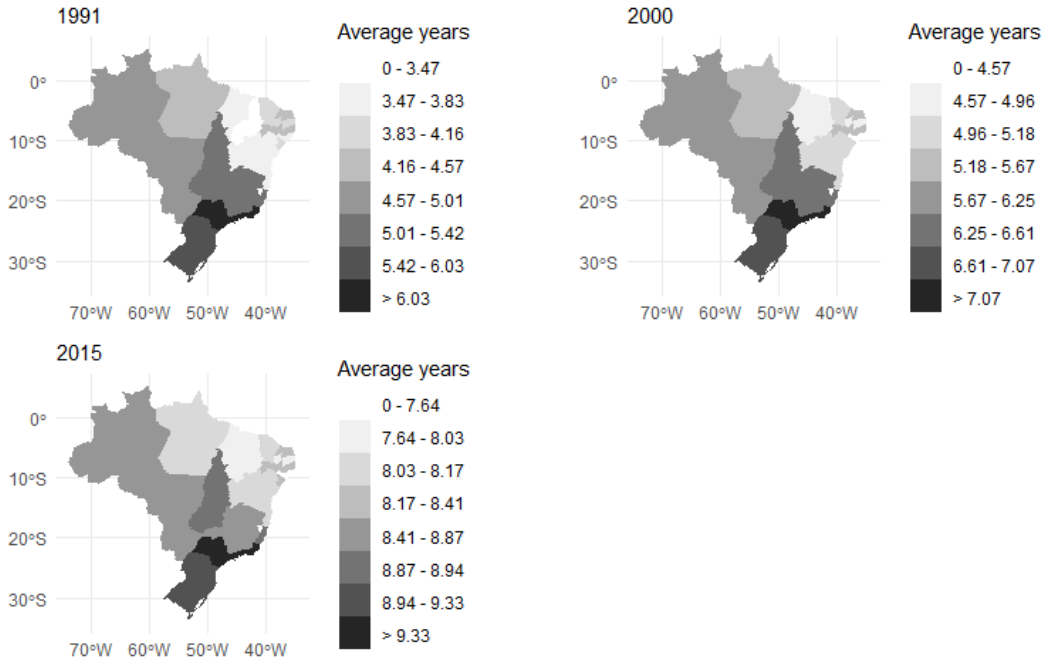
Figure 15: Average years of schooling by Brazilian state, population aged between 15 and 64 years, 1950-1980



Source: Own elaboration, see Section 3.

Note: We divided the map into percentiles: 15, 30, 45, 60, 75, 90, 100.

Figure 16: Average years of schooling by Brazilian state, population aged between 15 and 64 years, Brazil, 1991-2015

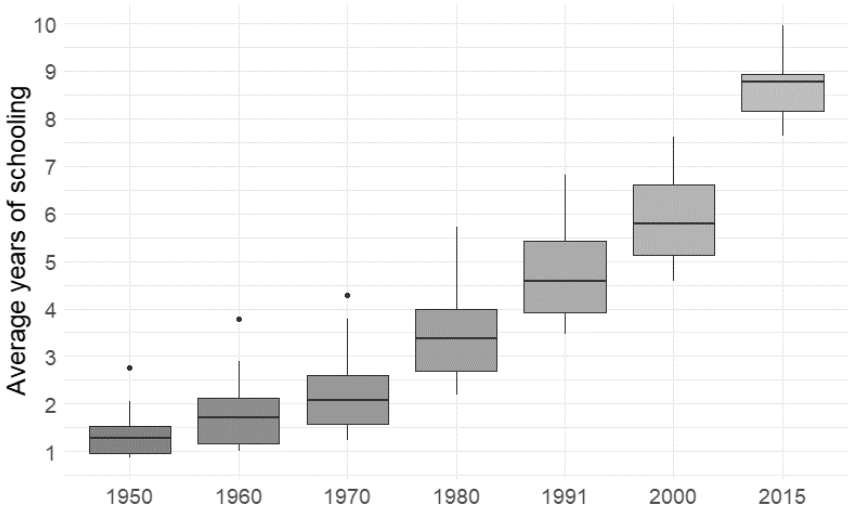


Source: Own elaboration, see Section 3.

Note: We have divided the map into percentiles: 15, 30, 45, 60, 75, 90, 100. Source: see Section 3.

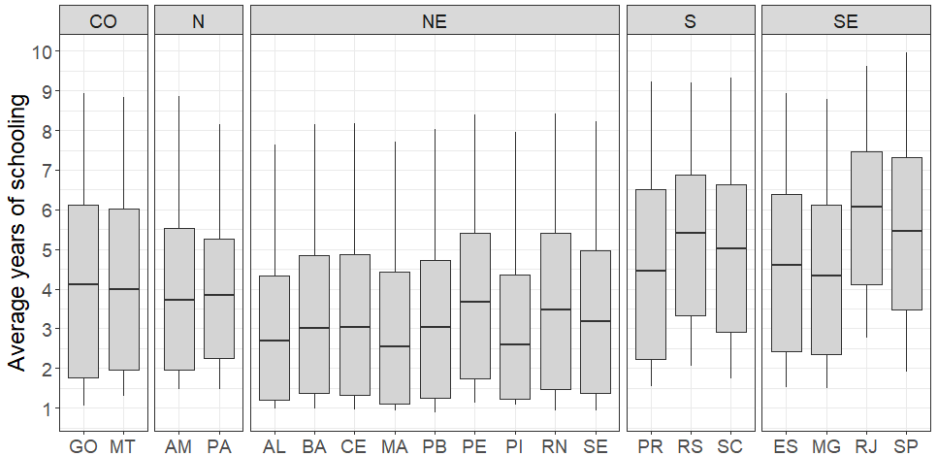
Figure 17 presents the dispersion of AYS from 1950 to 2015, by region. There was a high educational dispersion between states from 1960 to 2000, while some signs of convergence showed up from 2000 onwards. In addition, Figure 18 shows the dispersion by state from 1950 to 2015. It reinforces the conclusion that states that started from a more privileged position in 1950 tend to obtain better educational results (SE and S continue to be the most educated regions) in more recent periods (highest points represent the most recent period, the lowest correspond to the 1950 period).

Figure 17: Dispersion of average years of schooling, Brazil, 1950-2015



Source: Own elaboration, see Section 3.

Figure 18: Dispersion of average years of schooling by state, Brazil, 1950-2015



Source: Own elaboration, see Section 3.

Note: SE = Southeast; S = South; CO = Central-west; N = North; NE = Northeast.

5 Final remarks

This paper presented a novel annual series of average years of schooling (AYS) for Brazil from 1925 to 2015, which includes AYS information broken down by gender (male and female), race/color, and by states and regions. Contrary to the well-known series of Barro and Lee (2018), our results show that there was a gradual increase in AYS in Brazil between 1925 and 2015. Starting from 1.1 years in 1925,

Brazil reached 9.0 years in terms of AYS in 2015. Although there was a significant increase, the working-age population still has, on average, little more than the equivalent of a complete primary level.

We also observed a significant achievement in female education during the 1980s. In 1983, women surpassed men in terms of AYS. While men's AYS increased from 1.3 in 1925 to 8.7 in 2015, women's AYS rose from 0.9 in 1925 to 9.3 in 2015. In an analysis by racial groups, the yellow (Asian-Brazilian) population surpassed the whites in 1939 and has remained in the lead since then. However, the colored and black (Afro-Brazilian) groups have lagged behind, although there has been some slow convergence in recent years. Our dataset also covers regional information for the period 1950-2015. We found persistent educational inequality in regional terms: the Southeast region reported 2.0 years of schooling in 1950 and 9.6 in 2015, while the Northeast had 1.0 years of schooling in 1950 and reached 8.1 in 2015. The evolution of schooling in the Central-West region is also noteworthy: it started the series with 1.1 years in 1950 and ended with 8.9 years in 2015.

Our study has several implications for the international literature on AYS. While Barro and Lee (2013) established the research agenda on AYS, our paper reinforces the need for methodological changes in BL and derived datasets (e.g., Lee and Lee (2016)), as other studies have also emphasized. In the case of Brazil, it is advisable to avoid using the BL dataset, particularly given the implausible drop in AYS after 1970. Lutz et al. (2018) report results that are closest to our estimates, but their figures are questionable after 2010. Finally, it is clear that more studies are needed to provide more reliable information on educational attainment. Ideally, using specific methodologies for each country may deliver more reliable estimates even for comparative purposes, especially if there are studies for many countries. To facilitate further studies, a repository with material describing the construction of the dataset is available on GitHub.[35]

References

AGHION, P.; HOWITT, P. A model of growth through creative destruction. **Econometrica**, v. 60, n. 2, p. 323–351, 1992.

BAER, W. Regional inequality and economic growth in Brazil. **Economic Development and Cultural Change**, v. 12, n. 3, p. 268–285, 1964.

BARBOSA FILHO, F.; PESSÔA, S. A. Educação e crescimento: o que a evidência empírica e teórica mostra? **Economia**, v. 11, n. 2, p. 265–303, 2010.

BARBOSA, R. J. **Instruções para o uso dos bancos de microdados das amostras dos Censos**

[35] https://github.com/juliarwalter/pesquisa-anos-medios-de-escolaridade. Scripts for article graphs are also in the repository: https://github.com/juliarwalter/pesquisa-anos-medios-de-escolaridade-complemento.

Demográficos Brasileiros (1960 a 2010). Rio de JaneiroCentro de Estudos da Metrópole, , 2013.

BARRO, R. J. **Inequality, Growth, and InvestmentNBER Working Paper**. Cambridge, MA: National Bureau of Economic Research, mar. 1999. Disponível em: < http://www.nber.org/papers/w7038.pdf >.

BARRO, R. J.; LEE, J. W. International measures of schooling years and schooling quality. **American Economic Review**, v. 86, n. 2, p. 218–223, 1996.

BARRO, R. J.; LEE, J. W. International data on educational attainment: updates and implications. **Oxford Economic Papers**, v. 53, n. 3, p. 541–563, 2001.

BARRO, R. J.; LEE, J. W. A new data set of educational attainment in the world , 1950 – 2010. **Journal of Development Economics**, v. 104, p. 184–198, 2013.

BARRO, R. J.; LEE, J. W. **Barro-Lee Educational Attainment Dataset**. , 2018. Disponível em: < http://www.barrolee.com/ >

BARRO, R.; LEE, J. W. International comparisons of educational attainment. **Journal of Monetary Economics**, v. 32, n. 3, p. 363–394, dez. 1993.

BARROS, A. R. C. Dinâmica das desigualdades regionais no Brasil. **Revista Econômica do Nordeste**, v. 43, n. 4, p. 9–26, 2012.

BECKER, G. S. **Human capital: a theoretical and empirical analysis, with special reference to education**. first ed. New York: National Bureau of Economic Research, 1964.

COHEN, D.; LEKER, L. Health and Education: another look with the proper data. **CEPR Discussion Paper No. DP9940**, p. 1–25, 2014.

COHEN, D.; SOTO, M. Growth and Human Capital: good data, good results. **Journal of Economic Growth**, v. 12, n. 1, p. 51–76, mar. 2007.

DE CARVALHO FILHO, I.; MONASTERIO, L. Immigration and the origins of regional inequality: Government-sponsored European migration to southern Brazil before World War I. **Regional Science and Urban Economics**, v. 42, n. 5, p. 794–807, set. 2012.

DE CASTRO, M. H. G. **As desigualdades regionais no sistema educacional brasileiro**. Brasília: INEP, 1999.

DE LA FUENTE, A.; DOMÉNCH, R. Human capital in growth regressions: how much difference does data quality make? **Journal of the European Economic Association**, v. 4, n. 1, p. 1–36, 2006.

DE MELO, H. P.; THOMÉ, D. **Mulheres e poder: histórias, ideias e indicadores.** Rio de Janeiro: FGV Editora, 2018.

DENSLOW, D. **As Origens da Desigualdade Regional no Brasil**. São Paulo: [s.n.]. v. 3

EASTERLY, W.; LEVINE, R. Troubles with the Neighbours: Africa's Problem, Africa's Opportunity. **Journal of African Economies**, v. 7, n. 1, p. 120–42, 1998.

FÖLDVÁRI, P.; VAN LEEUWEN, B. Average Years of Education in Hungary: annual Estimates, 1920-2006. **Eastern European Economics**, v. 47, n. 2, p. 5–20, mar. 2009.

FÖLDVÁRI, P.; VAN LEEUWEN, B. Educational and income inequality in Europe, ca. 1870–2000. **Cliometrica**, v. 8, n. 3, p. 271–300, nov. 2013.

FRANKEMA, E. **The historical evolution of inequality in Latin America: a comparative analysis, 1870-2000**. PhD Thesis—[s.l.] Rijksuniversiteit Groningen, 2008.

GONZALEZ, C. A. G.; AIDAR, T. **Análise de pseudo-coortes a partir dos censos demográficos no brasil: uma aproximação metodológica**. CampinasTextos Nepo, 71, , 2015. Disponível em: < http://www.nepo.unicamp.br/publicacoes/textos_nepo/textos_nepo_71.pdf >

HALL, R. E.; JONES, C. I. Why do some countries produce so much more output per worker than others ? **The Quarterly Journal of Economics**, v. 114, n. 1, p. 83–116, 1999.

HANUSHEK, E. A.; WOESSMANN, L. Do better schools lead to more growth ? Cognitive skills , economic outcomes , and causation. **Journal of Economic Growth**, v. 17, n. 4, p. 267–321, 2012.

KANG, T. H. **Instituições, voz política e atraso educacional no Brasil, 1930-1964**. PhD Thesis—[s.l.] FEA-IPE-USP, 2010.

KANG, T. H. Education and development projects in Brazil, 1932-2004: a critique. **Brazilian Journal of Political Economy**, v. 38, p. 766–780, 2018.

KANG, T. H.; PAESE, L. H. Z.; FELIX, N. F. A. Late and unequal: measuring enrolments and retention in Brazilian education, 1933-2010. **Revista de Historia Económica / Journal of Iberian and Latin American Economic History**, v. 39, n. 2, p. 191–218, set. 2021.

KC, S. et al. Projection of populations by level of educational attainment, age, and sex for 120 countries for 2005-2050. **Demographic Research**, v. 22, p. 383–472, mar. 2010.

KYRIACOU, G. A. **Level and growth effects of human capital: a cross-country study of the convergence hypothesis**. C.V. Starr Center for Applied Economics, New York University, 1991. Disponível em: < https://econpapers.repec.org/RePEc:cvs:starer:91-26 >.

LAU, L. J.; LOUAT, F. F.; JAMISON, D. T. **Education and productivity in developing countries an aggregate production**. 1991.

LEE, J. W.; LEE, H. Human capital in the long run. **Journal of Development Economics**, v. 122, p. 147–169, set. 2016.

LEFF, N. H. Economic development and regional inequality: origins of the Brazilian. **The Quarterly Journal of Economics**, v. 86, n. 2, p. 243–262, 1972.

LINDERT, P. H. **Making Social Spending Work**. Cambridge: Cambridge University Press., , 2021.

LUCAS, R. E. On the mechanics of economic development. **Journal of Monetary Economics**, v. 22, n. 1, p. 3–42, jul. 1988.

LUTZ, W. et al. Reconstruction of populations by age , sex and level of educational attainment for 120 countries for 1970-2000. **Vienna Yearbook of Population Research 2007,** v. 5, n. 1, p. 193–235, 2007.

LUTZ, W. et al. **Demographic and Human Capital scenarios for the 21st century: 2018 assessment for 201 countries**. Luxembourg: Publications Office of the European Union, 2018.

MANKIW, N. G.; ROMER, D.; WEIL, D. N. A contribution to the empirics of economic growth. **The Quarterly Journal of Economics**, v. 107, n. 2, p. 407–437, maio 1992.

MINCER, J. Investment in Human Capital and personal income distribution. **Journal of Political Economy**, v. 66, n. 4, p. 281–302, ago. 1958.

MONASTERIO, L. M. Brazilian spatial dynamics in the long term (1872–2000): "path dependency" or "reversal of fortune"? **Journal of Geographical Systems**, v. 12, n. 1, p. 51–67, mar. 2010.

MORRISSON, C.; MURTIN, F. The century of education. **Journal of Human Capital**, v. 3, n. 1, p. 1–42, mar. 2009.

MULLIGAN, C.; SALA-I-MARTIN, X. **Measuring aggregate Human CapitalJournal of Economic Growth**. Cambridge, MA: National Bureau of Economic Research, fev. 1995. Disponível em: < http://www.nber.org/papers/w5016.pdf >.

NARITOMI, J.; SOARES, R. R.; ASSUNÇÃO, J. J. Institutional Development and Colonial Heritage within Brazil. **The Journal of Economic History**, v. 72, n. 2, p. 393–422, maio 2012.

NEHRU, V.; SWANSON, E.; DUBEY, A. A new database on Human Capital stock in developing and industrial countries: sources, methodology, and results. **Journal of Development Economics**, v. 46, n. 2, p. 379–401, abr. 1995.

OLIVEIRA, R. C.; SILVEIRA NETO, R. DA M. Expansão da escolaridade e redução da desigualdade regional de renda no Brasil entre 1995 e 2011: progressos recentes e desafios presentes. **Pesquisa e Planejamento Econômico**, v. 46, n. 1, p. 41–65, 2016.

PEREIRA, T. A. Z. Taxation and the stagnation of cotton exports in Brazil, 1800–60. **The Economic History Review**, v. 74, p. 522-545, 2021.

PESSÔA, S. A. **Existe um problema de desigualdade regional no Brasil?** Salvador, 2001.

PRITCHETT, L. Where has all the education gone? **The world bank economic review**, v. 15, n. 3, p. 367–391, 2001.

PSACHAROPOULOS, G. **The educational composition of the labor force: an international update**. World Bank, , 1992.

PSACHAROPOULOS, G.; ARRIAGADA, A. M. The educational composition of the labour force: an international comparison. **International Labour Review**, v. 125, n. 5, p. 561–574, 1986.

RAJAN, R.; ZINGALES, L. **Financial dependence and growthAmerican Economic Review**. Cambridge, MA: National Bureau of Economic Research, set. 1996. Disponível em: < http://www.nber.org/papers/w5758.pdf >.

RAMEY, G.; RAMEY, V. A. Cross-country evidence on the link between volatility and growth. **The American Economic Review**, v. 85, n. 5, p. 1138–51, 1995.

REIS, E. Spatial income inequality in Brazil, 1872–2000. **EconomiA**, v. 15, n. 2, p. 119–140, maio 2014.

RIBEIRO, S. C. A pedagogia da repetência. **Estudos Avançados**, v. 5, n. 12, p. 07–21, ago. 1991.

ROMER, P. M. Increasing Returns and Long-Run Growth Author. **The University of Chicago Press**, v. 94, n. 5, p. 1002–1037, 1986.

ROMER, P. M. Endogenous technological change. **The Journal of Political Economy**, v. 98, n. 5, p. S71–S102, 1990.

SACHS, J.; WARNER, A. **Natural resource abundance and economic growthNBER**. Cambridge, MA: National Bureau of Economic Research, dez. 1995. Disponível em: < http://www.nber.org/papers/w5398.pdf >.

SCHULTZ, T. W. Investment in Human Capital. **American Economic Association**, v. 51, n. 5, p. 1035–1039, 1961.

SOLOW, R. M. A contribution to the theory of economic growth. **The Quarterly Journal of Economics**, v. 70, n. 1, p. 65–94, 1956.

SPERINGER, M. et al. **Validation of the Wittgenstein centre back-projections for populations by age , sex , and six levels of education from 2010 to 1970**. LaxenburgIIASA Interim Report, , 2015.

UNESCO INSTITUTE FOR STATISTICS - UIS. **UIS Methodology for estimation of mean years of schooling**. UNESCO Institute for Statistics, , 2013.

UNITED NATIONS. **Convention on the elimination of all forms of discrimination against women**. Treaty Series, , 1988.

UNITED NATIONS DEVELOPMENT PROGRAMME - UNDP. **Human Development Report 2020: the next frontier human development and the anthropocene**. [s.l: s.n.].

VAN LEEUWEN, B.; VAN LEEUWEN-LI, J. Education since 1820. Em: **How Was Life?** [s.l.] OECD, 2014. p. 87–100.

VAN LEEUWEN, B.; VAN LEEUWEN-LI, J.; FÖLDVÁRI, P. **Average years of education (Average, total Population 15 years and older), 1850-2010**. , 2016. Disponível em: < https://clio-infra.eu/Indicators/AverageYearsofEducation.html >

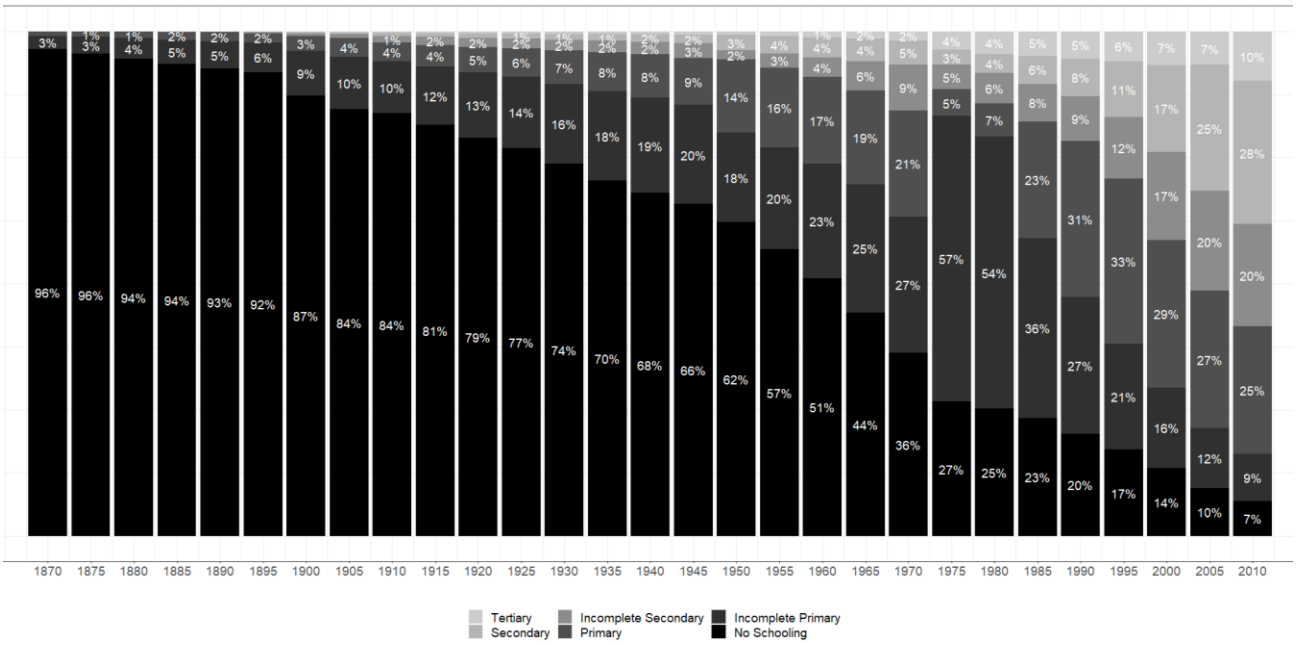
WILLIAMSON, J. G. Regional inequality and the process of national development: a description of the patterns. **Economic Development and Cultural Change**, v. 13, n. 4, p. 1–84, 1965.

WJUNISKI, B. S. Education and development projects in Brazil (1932-2004): political economy perspective. **Revista de Economia Política**, v. 33, n. 1, p. 146–165, mar. 2013.

WOESSMANN, L. Specifying Human Capital. **Journal of Economic Surveys**, v. 17, n. 3, p. 239–270, jul. 2003.

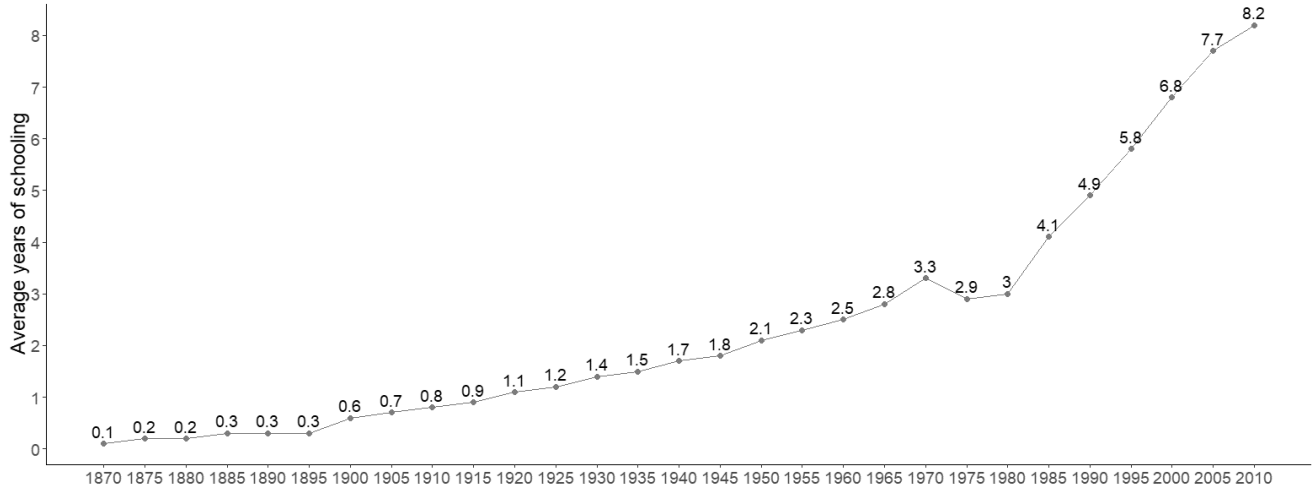
Appendix A - Lee and Lee Data (2016)

Figure A1: Educational attainment distribution, population aged 15 or over, Brazil (Lee and Lee 2016)



Source: Own elaboration based on Lee and Lee (2016) .

Figure A2: Average years of schooling, population aged between 15 and 64 years, Brazil (Lee and Lee 2016)



Source: Own elaboration based on Lee and Lee (2016).

Appendix B – Microdata

- **1960 Census**: These data are self-weighted and have a low sample fraction, around 1.25% of the population, allowing only a representative analysis for the states. This census collected information

about present, absent, and non-resident population. In the case of absent residents, the present ones provided the needed information. Non-residents (individuals who were visiting the household at the time of the interview) were then excluded from the analysis to avoid double-counting (Barbosa 2013). There are two questions of interest in the census used to calculate average years of schooling (AYS): the last grade completed with approval and the grade of the grade completed. People who attended or were attending school were supposed to answer these questions.

- **1970 Census**: It presents a representative sample of 25% (also for intra-municipal regions). As in 1960, non-residents were excluded and the questions applied to the population who attended/were attending school is similar. This census also contains information on people who are studying for entrance examinations and adult education and has the disadvantage of aggregating all individuals who attended the 5th and 6th grade (no matter the level) in the same variable, to which 17 years of study were attributed.[36]
- **1980 Census**: It also features a 25% sample of the population. In this year and subsequent years, only data from residents are collected (absent or present), that is, no prior filtering of information is necessary for the analysis (Barbosa 2013). From 1980 onwards, there is a change in the questionnaire in order to cover the education reform (Law 5692/1971). Different sets of questions are applied to individuals who were attending schools at the time of the interview and those who have ever attended schools at some point in the past. From this census on, the questionnaire also started to provide information on adult education through TV or radio.
- **1991 Census**: Sample of 25% of the population. Both the 1991 census and the 2000 census are similar to the 1980 census regarding educational data collection.
- **2000 Census**: Sample of 10% of the population.
- **2010 Census:** It was not possible to apply the same matching methodology, because the questionnaire does not contain any question regarding the specific grade for the population who has ever attended school in the past (Gonzalez and Aidar 2015). To overcome the problem, we used household survey data (PNAD-C)
- **PNAD Contínua, 2012 to 2015**: Data from the old PNAD were not used for two reasons: (i) the survey was discontinued and (ii) the presence of large sample fluctuations. PNAD-C, the new household survey, has a larger sampling plan (less subject to fluctuations). Therefore we use PNAD-C to estimate schooling levels in recent years. To estimate the educational distribution, we took information from the first quarter of each year.

[36] The proportion of students in the tertiary level in the 5th and 6th grade was much bigger than the others levels, since graduate students were also assigned in this category.

We also used the following criteria to assign years of schooling in some unclear cases:

a) **individuals who attended more grades than what was usually required to complete a schooling level:** extra years of schooling were not added. In 1960, for example, those who completed the 5th grade of the old primary education were assigned four years of schooling instead of five, since four grades are the standard duration and there is no way to know the number of years attended in several cases (for individuals who attended more advanced levels of education, we do not have information on the number of grades completed in primary schools). The same procedure was adopted for the other levels of education. In the case of PNA-C, Law 11.274/2006 extended the length of the new primary education to nine years. Those who completed the 9th grade were assigned eight years of schooling. The latter issue is a minor one since students who entered the nine-year primary education in 2007 had not reached the 9th grade in 2015, the final year of our dataset;

b) **adult literacy and non-serial education:** we excluded individuals from the dataset when it was not possible to determine the last grade attended; exceptions were students enrolled in preparatory courses and master's students, who were assigned 11 and 17 years respectively;

c) **upper bound of years of schooling:** we set a maximum limit of 17 years of schooling. This standardization was carried out to avoid distortions when comparing census data with information from PNAD-C.