Universidade de São Paulo

Faculdade de Economia, Administração e Contabilidade

Departamento de Economia

Programa de Pós-Graduação em Economia

O que o Nome nos ensina? Padrões sociais e raciais de nomes e sobrenomes e performance escolar em São Paulo.

Lucas Costa Scottini

Orientador: Prof. Dr. Marcos de Almeida Rangel

Outubro de 2011

São Paulo

Prof. Dr. João Grandino Rodas
Reitor da Universidade de São Paulo

Prof. Dr. Reinaldo Guerreiro
Diretor da Faculdade de Economia, Administração e Contabilidade

Prof. Dr. Denisard Cnéio de Oliveira Alves
Chefe do Departamento de Economia

Prof. Dr. Pedro Garcia Duarte
Coordenador do Programa de Pós-Graduação em Economia

Lucas Costa Scottini

O que o Nome nos ensina? Padrões sociais e raciais de nomes e sobrenomes e performance escolar em São Paulo.

Dissertação apresentada ao Departamento de Economia da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo como requisito para a obtenção do título de Mestre em Economia.

Orientador: Prof. Dr. Marcos de Almeida Rangel

Outubro de 2011

Versão Original

São Paulo

FICHA CATALOGRÁFICA

Elaborada pela Seção de Processamento Técnico SBD/FEA-USP

Scottini, Lucas Costa

O que o Nome nos ensina? Padrões sociais e raciais de nomes e sobrenomes e performance escolar em São Paulo / Lucas Costa Scottini - São Paulo, 2011.

82 p.

Dissertação (Mestrado) - Universidade de São Paulo, 2011

Orientador: Marcos de Almeida Rangel.

1. Educação - Aspectos Econômicos 2. Econometria 3. Discriminação I. Universidade de São Paulo. Faculdade de Economia, Administração e Contabilidade. II. Título.

CDD - 370.193

À minha família

AGRADECIMENTOS

Agradeço, em primeiro lugar, aos meus pais, Sandra e Dionísio, pela oportunidade da vida e por todos os incentivos que recebi durante minha trajetória acadêmica. Esta dissertação, sem dúvidas, é muito mais deles do que minha propriamente. Reconheço também o apoio dos meus irmãos Leonardo e Leandra.

A todos os meus amigos de São Lourenço e São Paulo dedico um muito obrigado. Pela convivência e camaradagem, agradeço ao Gustavo Fasolo, Phelipe Eduardo e Ricardo Sabbadini. Agradeço também a Heitor Pellegrina, Thomaz Gemignani, Dejanir Silva e a todos os outros colegas de mestrado do IPE pelo excelente ambiente e pelas inúmeras discussões de idéias.

Agradeço ainda ao meu orientador Marcos Rangel e ao meu co-orientador Ricardo Madeira pelas sugestões e pela liberdade criativa. Outros professores que de alguma forma contribuíram para este trabalho foram Fernando Botelho, Mauro Rodrigues, Dante Aldrighi e Ana Maria Bianchi. Meu muito obrigado a eles também.

Por último, agradeço à FAPESP, pelo apoio financeiro.

- O meu nome é Severino,
não tenho outro de pia.
Como há muitos Severinos,
que é santo de romaria,
deram então de me chamar
Severino de Maria;
(...)
Somos muitos Severinos
iguais em tudo na vida:
na mesma cabeça grande
que a custo é que se equilibra,
no mesmo ventre crescido
sobre as mesmas pernas finas,
e iguais também porque o sangue
que usamos tem pouca tinta.
João Cabral de Mello Neto

RESUMO

Esta dissertação estuda padrões sociais e raciais de nomes e sobrenomes entre alunos paulistas e a correlação desses padrões com desempenho escolar. Para cada nome e sobrenome observados criamos índices que medem o quão distintamente rico (ou pobre) e o quão distintamente branco (ou afro) tais nomes e sobrenomes são. Os resultados são de que tanto nomes quanto sobrenomes dão sinais de status socioeconômico, enquanto sobrenomes e, em menor medida, nomes predizem raça. Nomes que indicam maior status são, em geral, de maior frequência estatística, de origem latina, formados por um só termo e com grafia condizente com a Língua Portuguesa formal. Nomes de menor frequência estatística, compostos por duas palavras, com influência do idioma inglês na grafia e na pronúncia e com grafias distintas do Português formal estão associados a baixo status. A evidência aponta ainda para uma relação mais forte entre primeiro nome e status do que entre primeiro nome e raça, revelando que, em São Paulo, o universo cultural que baseia a escolha de primeiros nomes é mais classe-específico que raça-específico. Pelo lado dos nomes de família, os dados mostram que sobrenomes tipicamente de alto status e tipicamente brancos são de menor frequência estatística e de origem não-portuguesa. Sobrenomes portugueses não apresentam padrão socioeconômico e racial, com exceção dos três sobrenomes mais frequentes, os quais são tipicamente pobres. Além disso, sobrenomes associados à religião católica são tipicamente pobres e afrodescendentes. Posteriormente, avaliamos a relação entre diversas medidas de performance escolar e nomes e sobrenomes. Os dados apontam para uma associação robusta entre ter um nome distintamente pobre (rico) e um sobrenome distintamente afro (branco) e piores (melhores) resultados escolares. Tal evidência é consistente com um cenário em que escolhas culturais dos pais e herança cultural familiar afetam a acumulação de capital humano de crianças e jovens adultos. Entre outras possibilidades está a existência de tratamento discriminatório nas escolas.

ABSTRACT

This dissertation studies social and racial patterns of names and surnames for students in São Paulo and evaluates its contribution to academic performance. For each name and surname observed, we create indices to measure how distinctively rich (or poor) and how distinctively white (or black) they are. We find that both names and surnames are predictors of social status, while surnames and, to a lesser extent, names predict race. Names that indicate higher status present, generally, higher frequency, latin origin, only one term and spelling coherent to formal Portuguese language. Names that signal low status show lower frequency, two terms, English language influence on spelling and pronouncing and spelling different from formal Portuguese. The evidence also points to a stronger link between first names and social status than between first names and race, revealing that cultural determinants of name choice are mainly social rather than racial. When it comes to family names, data shows that distinctively high status and white surnames have low frequency and non-portuguese origin. Portuguese surnames do not present social or racial patterns, except for the three most common surnames, which are typically poor. Additionaly, catholic devotional surnames are distinctively poor and afro. Then we evaluate the correlation between names, surnames and a set of academic performance measures. The evidence indicates a robust link between having a distinctively poor (rich) name and a distinctively afro (white) family name and worse (better) grades. This is consistent with a scenario where parental cultural choices and familiar cultural heritage affect human capital accumulation of children. Discriminatory treatment in school is another possibility.

Sumário

1. Introdução 3
2. Descrição dos Dados 12
   - 2..1 Base de Matrículas 12
   - 2..2 Saresp 2008 15
   - 2..3 Notas do Professor 16
3. Sobre os índices sociais e raciais de nomes e sobrenomes 17
   - 3..1 Indicador socioeconômico 17
   - 3..2 Indicador Racial 20
4. Padrões Sociais e Raciais de Nomes e Sobrenomes 22
   - 4..1 Primeiros Nomes 22
   - 4..2 Sobrenomes ou Nomes de Família 37
5. Nomes, Sobrenomes e Desempenho Escolar 49
6. Conclusão 55
7. Apêndice 58
8. Referências Bibliográficas 65

Lista de Tabelas

1 Distribuição da Quantidade de Palavras no Nome Completo 13

2 Matriz de correlações dos Índices 20

3 Matriz de correlações dos rankings dos Índices 20

4 Nomes mais Frequentes 22

5 Nomes de Meninos mais frequentes por tipo de Escola 23

6 Nomes de Meninas mais frequentes por tipo de Escola 24

7 Nomes com maiores ISN para Meninos 29

8 Nomes com menores ISN para Meninos 31

9 Nomes com maiores ISN para Meninas 32

10 Nomes com menores ISN para Meninas 33

11 Determinantes do ISN e do IRN 34

12 Regressões do ISN contra características ortográficas dos Nomes 35

13 Nomes mais frequentes terminados no sufixo SON 37

14 Sobrenomes mais Frequentes 39

15 Média do ISS por origem 44

16 Patronímicos Portugueses e Espanhóis 45

17 Sobrenomes mais frequentes por origem 46

18 Letras Estrangeiras no Sobrenome e ISS 46

19 Sobrenomes Portugueses mais Frequentes e seus Índices 47

20 Sobrenomes Devocionais e seus Índices 48

21 Proficiência em Português no exame SARESP 2008 58

22 Proficiência em Matemática no exame SARESP 2008 59

23 Proficiência em Ciências no exame SARESP 2008 59

24 Nota média no boletim de Português de 2008 60

25 Nota média no boletim de Matemática de 2008 60

26 Nota média no boletim de Ciências de 2008 61

27 Nota média no boletim de Educação Física de 2008 61

28 Nota média no boletim de Artes de 2008 62

29 Estatísticas Descritivas 63

1. Introdução

Qual é a primeira pergunta de um questionário ou a primeira linha de qualquer currículo profissional? Nomes e sobrenomes são os principais identificadores de um indivíduo em sua vida social e contribuem de maneira importante na formação da imagem que o indivíduo tem de si mesmo. Mas, à parte disso, o quanto podemos aprender sobre a origem familiar de alguém através de seu nome e sobrenome? Em geral, esperamos que nomes sejam indicadores do universo cultural dos pais e das expectativas que esses embutem em seus filhos. Afinal de contas, todos os anos milhões de pais ao redor do mundo se defrontam com o desafio de escolher o nome certo para seus filhos, e o fazem lastreados por suas preferências culturais. Podemos pensar que o nome reflete a cultura que os pais decidiram adotar. Por outro lado, sobrenomes indicam outras dimensões da família de um indivíduo. Sobrenomes não refletem uma escolha dos pais, eles são herdados das gerações anteriores. Em especial, sobrenomes têm a dizer sobre a origem imigratória de um indivíduo e, em muito casos, a posição de sua família na hierarquia social. Enquanto o nome reflete a cultura que os pais resolveram adotar, o sobrenome revela a cultura herdada.

Uma recente literatura em Economia tem revelado dois fenômenos importantes a respeito de nomes e sobrenomes. O primeiro deles é o de que nomes possuem padrão social e racial. Em outras palavras, em determinados contextos os diferenciais raciais e sociais entre indivíduos evoluíram para diferenças culturais, e tais diferenças culturais se manifestam quando da escolha dos nomes dos filhos. O segundo fenômeno relevante é o de que, além de serem indicadores de origem social e raça, nomes e sobrenomes parecem influenciar resultados socioeconômicos dos indivíduos. Os casos mais emblemáticos estudados pela literatura se referem a nomes típicos de afrodescendentes nos EUA e nomes e sobrenomes africanos e árabes na Europa Ocidental.

Fryer e Levitt (2004), usando dados de nascimentos no estado da Califórnia entre 1961/2000,

revelam que a partir do final dos anos 60 pais afrodescendentes norte-americanos passaram a escolher nomes nitidamente diferentes daqueles escolhidos por pais brancos. Segundo os autores, o movimento de orgulho negro que marcou os anos 60 inaugurou uma era em que os negros norte-americanos passaram a reafirmar sua identidade racial através da escolha de nomes tipicamente negros para seus filhos. Ainda há evidência de que mães solteiras, jovens, pouco educadas e com nomes distintamente negros têm maiores chances de escolher um nome tipicamente negro para os filhos.

Uma pergunta natural é a de se indivíduos com nomes tipicamente negros pagam algum custo por conta disso. Betrand and Mullainatham (2004) usam um experimento aleatório para descobrir se currículos de trabalhadores com nomes tipicamente negros ou brancos têm as mesmas chances de serem chamados para uma entrevista de emprego nas cidades de Boston e Chicago. A conclusão é a de que, em média, um currículo de nome negro precisa ser enviado 15 vezes para receber um convite de entrevista, contra 10 do currículo com nome branco. Além disso, o aumento das qualificações de um currículo não se reflete em maiores convites de entrevista para currículos com nomes tipicamente negros. Tais resultados não diferem substancialmente entre empregadores de diferentes tamanhos e setores de atuação, existindo até mesmo entre empresas que se declaram "equal opportunity employers".

Arai e Thoursie (2009) apresentam evidência para o mercado de trabalho sueco de que imigrantes que mudam seus sobrenomes originais para sobrenomes que soam suecos experimentam, em média, um acréscimo de 26% no salário. Em geral tais sobrenomes originais são de origem árabe (sendo Mohammed o caso mais simbólico) e eslava (como Ibrahimovic). O efeito da mudança de sobrenomes é mais acentuado na cauda inferior da distribuição de salários e há evidência de relação entre sobrenome e duração do desemprego. Arai et. al (2009) mostram que imigrantes na França, especialmente os provenientes da região do Magrebe e Oriente Médio, batizam seus filhos com nomes nitidamente difer-

entes dos nomes franceses convencionais. Contudo, pais nascidos na França, mas que são descendentes de imigrantes, escolhem nomes tipicamente franceses para seus filhos. Tais trabalhos revelam os esforços dos imigrantes na Europa para assimilar a cultura local através das escolhas de nomes e sobrenomes.

Na Espanha, ter um sobrenome mais incomum está correlacionado a maior status socioeconômico, como mostram Collado, Ortin e Romeu (2008). Segundo os autores, no passado, certas famílias tornavam seus sobrenomes mais raros ao unir em um único sobrenome hifenizado dois sobrenomes comuns. Isso serviria como um legado de famílias bem-sucedidas para suas futuras gerações. Mesmo passadas várias gerações, os autores mostram que os sobrenomes hifenizados são mais frequentes entre indivíduos cujas profissões possuem maior status social (médicos, advogados e empresários) relativamente ao restante da população.

Outros artigos avaliam ainda o papel dos nomes em contextos educacionais. Figlio (2004), usando dados de um distrito escolar da Flórida, EUA, identifica as características de um nome que são capazes de predizer a educação da mãe. Sua análise conclui que, comparando pares de irmãos, alunos com nomes que soam como dados por mães mais educadas obtêm melhores resultados escolares, tanto em exames padronizados corrigidos anonimamente quanto em medidas de desempenho sujeitas à discricionariedade dos professores. Figlio (2004) mostra que meninos cujos nomes são mais frequentes entre meninas são menos disciplinados na escola.

Para o Brasil, a questão dos padrões raciais e sociais de nomes e sobrenomes e o papel destes sobre resultados individuais ainda foram pouco explorados. Os estudos brasileiros existentes em historiografia da escravidão e genética populacional se concentram basicamente em analisar a ligação entre sobrenomes dos indivíduos e ascendência africana.

Segundo o Dicionário da Escravidão negra no Brasil, de Clóvis Moura:

> O escravo africano e os seus descendentes não tinham sobrenome de família. Ou usavam o do grupo étnico do qual eram originários, como, por exemplo, Manuel Mina ou Pedro Angola, ou usavam o sobrenome do seu senhor por aquiescência deste. Atribui-se essa falta à brutalidade com que os africanos eram apresados pelos captores e traficantes, desfazendo completamente os grupos familiares. Assim, o escravo negro era um ser sem ancestralidade e não conseguia situar-se em termos de linhagem. Com isso, os senhores lograram, em parte, desfazer os vínculos tribais, mas, ao mesmo ensejaram sua união "ante a desgraça comum", na expressão do Conde dos Arcos.

Assumir o sobrenome português do senhor, contudo, não nos conta toda a história sobre sobrenomes e ex-escravos. Barbosa et al (1997) apresentam evidência de que, no Brasil, existe correlação entre ter um sobrenome devocional católico e ancestralidade negra. O cromossomo Y de homens negros com sobrenomes devocionais apresentam mais características negróides do que o de homens negros sem tais sobrenomes e o de homens brancos. Segundo Azevedo (1980), no Estado brasileiro da Bahia, durante os séculos XVIII e XIX, 90% dos negros livres não possuíam sobrenomes. Entre os negros que adotaram um sobrenome, 44% no século XVIII e 69% no século XIX escolheram sobrenomes devocionais católicos. Tais evidências indicam que ter um sobrenome católico devocional está ligado a mais ascendência africana. Isso é relevante pois um dos efeitos da miscigenação é atenuar a correlação entre cor da pele e origem afro. Portanto, mesmo entre indivíduos brancos, aqueles com sobrenomes devocionais teriam mais origem negra do que aqueles com outros tipos de sobrenome.

Nosso trabalho estuda os status socioeconômico e racial de primeiros nomes e nomes de família e como tais status afetam o desempenho escolar de alunos do ensino fundamental e médio no estado de São Paulo. Nossas inovações para a literatura supracitada estão as-

sociadas justamente ao fato de utilizarmos tanto nomes quanto sobrenomes, de fazermos caracterizações raciais e sociais destes e de trabalharmos no contexto brasileiro.

Se nomes refletem a cultura escolhida pelos pais e sobrenomes a cultura que estes herdaram dos antepassados, analisar conjuntamente primeiros nomes e nomes de família nos dá uma medida da importância relativa entre essas duas culturas sobre a acumulação de capital humano dos indivíduos. Adicionalmente, podemos avaliar como a dimensão social e racial do sobrenome impacta na escolha dos primeiros nomes.

Caracterizar primeiros nomes nas dimensões racial e social nos permite verificar qual a dimensão mais relevante na escolha do nome. Ao contrário do EUA, o Brasil pós-escravista não apresentou instituições formais de segregação racial e possui taxas de miscigenação superiores às taxas norte-americanas. Isso nos leva a crer que a dimensão relevante para a construção de diferenças culturais entre os indivíduos no Brasil seja a socioeconômica e não a racial. Em outras palavras, enquanto nos EUA a segregação racial produziu culturas distintas, as quais se manifestaram na escolha dos primeiros nomes, no Brasil a desigualdade social parece ser o fator crucial que gera as diferenças culturais e, consequentemente, a diferença entre nomes. Enquanto nos EUA desenvolveu-se a categoria de "nomes de negro", no Brasil a categoria primordial seria a de "nomes de pobre".

Pelo lado dos nomes de família, o estudo de seus padrões socioeconômicos e raciais nos abre uma série de possibilidades. Podemos estudar, por exemplo, o status socioeconômico e racial atual dos sobrenomes de imigrantes europeus e asiáticos que chegaram em massa ao Estado de São Paulo durante o final do século XIX e começo do século XX. Por outro lado, podemos analisar também o atual status social e racial dos sobrenomes ligados a ex-escravos. Além de saber o status atual de um sobrenome, temos condições de investigar se tal status joga algum papel na acumulação de capital humano das atuais gerações. Isso

contribui para o entendimento da transmissão intergeracional de status socioeconômico no Brasil.

Por fim, nossa última contribuição deriva do contexto em que trabalhamos. Por tratarmos de Brasil, um país em desenvolvimento e de língua portuguesa, podemos estudar qual é o padrão da influência da cultura de massa norte-americana nas escolhas dos primeiros nomes. A evidência anedótica é a de que a exposição à tal cultura, através do cinema e televisão, leva os brasileiros, em especial os mais pobres e menos educados, a adotar nomes que soam ingleses. Podemos avaliar empiricamente tal questão e testar se nomes de cunho inglês são sinais de baixo status socioeconômico no Brasil.

A partir da base de matrículas da Secretaria de Educação do Estado de São Paulo, que conta com 12.270.797 de estudantes, construímos um índice socioeconômico para cada nome e cada sobrenome encontrados. Tal índice se baseia nas frequências dos nomes e sobrenomes entre estudantes de escolas públicas e particulares, os quais assumimos serem estudantes de baixo/médio e alto status socioeconômico, respectivamente. Usando dados provenientes do SARESP, que contabiliza 3.111.128 observações, construímos um índice racial para cada nome e sobrenome encontrados, baseado nas suas frequências relativas entre estudantes afrodescendentes (negros e mulatos) e brancos. Nossa análise dos índices nos desvenda um rico conjunto de fatos.

Nomes associados a alto status social tendem a ter grafia coerente com a língua portuguesa formal, possuir apenas um termo (nomes simples) e apresentar maior frequência estatística. Ter maior frequência significa que a população de maior status social possui uma diversidade menor de primeiros nomes ou se concentra mais em poucos nomes. Exemplos de nomes de altíssimo status são Luisa, Sofia e Clara para meninas e Bernardo, Pedro e Heitor para meninos. Temos ainda que um parcela dos pais de alto status social escolhe

primeiros nomes de origem não-portuguesa, como: Matteo, Kenzo, Ibrahim e Otto para meninos; Martina, Sophie, Pietra e Beatrice para meninas.

Os nomes de baixo status social, por sua vez, tendem a ter inspiração na língua inglesa (revelando possivelmente a influência da cultura de massa norte-americana), são grafados com letras e sílabas que não constam do português oficial, são fusões entre os nomes da mãe e do pai, fazem referência a personagens do cristianismo, estão associados à cultura nordestina, possuem dois termos (nomes compostos) e menor frequência estatística - grande diversidade de nomes cada qual com pouca frequência estatística. Exemplos de nomes de baixo status são Jéssica Aparecida, Daiane, Leidiane e Raimunda para meninas, e Michael Douglas, Jackson, Josimar e Severino para meninos.

Primeiros nomes muito comuns, em geral, não apresentam correlação com status socioeconômico. Em geral, tais nomes possuem grafia coerente com a língua portuguesa e, especialmente para meninos, estão presentes no novo testamento da bíblia. São os casos de Lucas, Gabriel e Rafael para meninos e Jéssica, Beatriz e Letícia para meninas.

Filhos de mães e pais com ensino superior e com sobrenomes de maiores índices social e racial têm em média nomes de maior status que crianças de pais pouco educados e com sobrenomes de menores índices social e racial. Controlando por educação dos pais e pelos índices de sobrenome, não há diferença de status social nos nomes de crianças brancas e de crianças afrodescendentes.

O índice de status racial do nome apresenta uma concentração relevante em valores segundo os quais os nomes estão igualmente distribuídos entre crianças afrodescendentes e brancas. Em outras palavras, nomes raça-específico compõem uma minoria da população estudada, confirmando que nomes classe-específicos são muito mais pervasivos. Entre

nomes tipicamente brancos temos Theo, Gianlucca e Tarik para meninos e Nina, Chiara e Lia para meninas. Entre nomes tipicamente negros estão Oséias, Aldair e Washington para meninos e Alcione, Geane e Valdinéia para meninas.

Ao contrário dos nomes, sobrenomes mais incomuns são mais típicos entre crianças de maior status e brancas. Em geral, tais sobrenomes são de origem não-portuguesa. Sobrenomes tipiciamente portugueses (colonização majoritária), em geral, não apresentam padrão socioeconômico e racial. Tais sobrenomes tem alta frequência e estão democraticamente distribuídos na população. Por outro lado, os dois sobrenomes mais comuns da amostra, Silva e Santos, estão associados a baixo status e, no caso do segundo, à população afro. Sobrenomes adotados por ex-escravos, os chamados sobrenomes católico devocionais, são tipicamente pobres e negros. São os casos de sobrenomes como de Jesus, Bispo, Batista, dos Anjos, Nascimento e vários títulos de Maria (Nossa Senhora).

De posse dos índices raciais e sociais de nomes e sobrenomes, investigamos sua relação com a performance escolar de alunos do ensino fundamental e médio das escolas públicas estaduais do estado de São Paulo. A análise dos resultados de testes padronizados corrigidos sem a interferência humana (doravante, testes blind) nos revela que crianças de nomes tipicamente pobres (ricos) e sobrenomes tipicamente negros (brancos) apresentam pior (melhor) desempenho em Língua Portuguesa, Matemática e Ciências. Uma elevação conjunta de um desvio-padrão em cada uma das medidas de nome e sobrenome consideradas está associada a um aumento de .065 desvios-padrão no resultado em Português, .062 em Matemática e .065 em Ciências. Isso equivale a cerca de um terço da diferença de desempenho entre filhos de mães com ensino superior e mães sem educação formal, ou ainda a cerca de metade do gap entre alunos brancos e alunos negros. Os resultados se mantém quando medimos performance acadêmica através das notas do boletim escolar.

Há evidência ainda de que o efeito do nome é mais importante para o desempenho em Português do que em Matemática. Isso é consistente com um cenário em que nomes estão mais associados a habilidades linguísticas e hábitos de leitura dos pais. Outra regularidade apresentada pelos dados é a de que as medidas do sobrenome herdado do pai importam mais para a performance do aluno do que as medidas do sobrenome herdado da mãe.

Embora a principal interpretação seja a de que nomes e sobrenomes estejam correlacionados com características culturais não observadas que afetam desempenho escolar, não pode-se descartar um cenário em que exista tratamento discriminatório dentro das escolas públicas. Tal tratamento pode ser desencadeado pelas expectativas dos próprios professores. Se o que se espera de um aluno com nome pobre é ter pior desempenho que um aluno de nome rico, pode-se estabelecer um equilíbrio em que tais diferenciais esperados se tornem reais.

O restante desta dissertação está organizado da seguinte forma. Na terceira seção, descrevemos os bancos de dados utilizados na construção dos índices socieconômico e racial de nomes e sobrenomes e nas regressões de desempenho escolar. A construção dos índices é detalhada na quarta seção. Na quinta seção, fazemos uma descrição dos nomes, sobrenomes e seus padrões sociais e raciais. Na sexta, apresentamos os impactos dos nomes e sobrenomes sobre as medidas de desempenho escolar e discutimos os resultados. Por fim, guardamos a última seção para a conclusão.

2. Descrição dos Dados

2..1 Base de Matrículas

O banco de dados principal utilizado para a construção do índice socioeconômico dos nomes e sobrenomes constitui do cadastro de matrículas escolares do Estado de São Paulo para os anos de 2007, 2008, 2009 e 2010. Constam nessa base todos os alunos matriculados na rede pública municipal, estadual e federal e na rede privada de ensino infantil, fundamental e médio. Ao todo, a amostra contém 12.270.797 alunos. Como nossa estratégia é gerar um índice socioecônomico para cada nome e sobrenome, o tamanho da base é crucial. Para cada estudante, as informações disponíveis são: nome completo, gênero, data de nascimento, rede de ensino (pública ou particular) e nome completo dos pais. Considerando que as taxas de matrícula de crianças entre 7 e 14 anos no Estado de São Paulo superam 98%, temos uma base muito próxima do nível censitário para os nascidos entre os anos de 1993 e 2003.

As bases auxiliares são duas. A primeira consiste, para o ano de 2008, do cadastro dos estudantes paulistas beneficiários do Programa Bolsa Família (doravante, PBF), um programa de transferência de renda de abrangência nacional voltado para famílias carentes. Tal base possui 8.512.670 alunos (todos os matriculados no sistema público ou privado no Estado), dos quais 1.517.775 são beneficiários do PBF. A segunda constitui dos questionários socioeconômicos do Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (SARESP) para os anos de 2005, 2007 e 2008. Tais questionários formam um banco de dados com nome, sobrenome e raça dos alunos bem como o nível educacional do pai e da mãe, totalizando 3.111.128 estudantes da rede pública estadual do estado de São Paulo. Em suma, possuímos uma base principal e duas amostras menores, com informações adicionais de participação no PBF, educação dos pais e raça respectivamente. Esses dados serão utilizados para caracterizar tanto padrões socioeconômicos quanto raci-

ais de nomes e sobrenomes.

O primeiro passo de nossa análise foi o de contar o número de palavras que constituem um nome completo. 99,7% dos alunos possuem nome completo formado por 7 ou menos palavras. Para padronizar a análise, os 0,3% restantes foram excluídos da base. A seguir, dispomos uma tabela com a distribuição do número de palavras do nome completo.

Tabela 1: Distribuição da Quantidade de Palavras no Nome Completo

Número de Palavras	Contagem	Frequência (%)	Acumulado (%)
1	3	0	0
2	410.404	3,34	3,34
3	3.846.103	31,34	34,69
4	5.511.877	44,92	79,61
5	2.231.308	18,18	97,79
6	247.400	2,02	99,81
7	23.702	0,19	100

A tabela acima revela que a moda são nomes completos formados por 4 termos, seguido de 3 e 5 termos. Além disso, nomes formados por somente dois termos, uma palavra para o nome e uma palavra para o sobrenome, são minoria (apenas 3,34% da amostra). O costume no Brasil é a criança assumir dois sobrenomes: o primeiro sendo o último sobrenome da mãe e o segundo sendo o último sobrenome do pai. Portanto, a transmissão de sobrenomes no Brasil é patrilinear, ou seja, somente o nome do pai sobrevive na segunda geração da família.

Posteriormente, excluímos do nome completo algumas palavras que precedem certos sobrenomes, são elas: al, el, dall, da, das, dal, de, dei, del, des, der, della, di, do, dos, du, dus, von, van, la, le, les, los, e. Por exemplo, o nome Edson Arantes do Nascimento, após a aplicação do critério de exclusão, ficaria Edson Arantes Nascimento. Luis Inácio Lula da Silva se tornaria Luis Inácio Lula Silva e assim por diante. Com isso, não iremos diferenciar os sobrenomes Oliveira e de Oliveira, por exemplo.

Dado que no banco de dados original não há distinção entre quais palavras formam o nome e quais formam o sobrenome, há a necessidade de se adotar um critério de separação de palavras. O critério mais simples seria o de considerar a primeira palavra do nome completo como sendo o nome, e todas as palavras restantes formariam o sobrenome. Apesar de sua praticidade, tal critério desconsidera a existência de nomes compostos, isto é, nomes formados por mais de uma palavra. Por exemplo, usando o critério da primeira palavra o nome Fernando Henrique Cardoso seria formado por um nome (Fernando) e dois sobrenomes (Henrique e Cardoso), quando na realidade a separação convencional é a que considera Fernando Henrique como nome e Cardoso como sobrenome ou nome de família. Dado que a prática de batizar os filhos com nomes compostos é comum no Brasil, e, mais importante, pode estar correlacionada com características socioeconômicas dos pais, precisamos de um critério que consiga abarcar tais nomes.

O nosso critério de separação entre nome e sobrenome se baseia em duas hipóteses: a de que todas as pessoas possuem pelo menos um nome e um sobrenome e que os nomes podem ser compostos de uma ou duas palavras (formando um nome simples e um nome composto, respectivamente). Uma implicação imediata dessas hipóteses é a de que pessoas cujo nome completo é formado por somente duas palavras sempre têm nome simples. Alternativamente, uma pessoa só pode deter um nome composto se seu nome completo for formado por, no mínimo, 3 palavras.

Uma pessoa do gênero masculino (feminino) terá nome composto, segundo nosso critério, se, e somente se, tiver um nome completo formado por no mínimo 3 termos e o segundo termo aparecer mais de 100 vezes na nossa base como primeiro termo do nome completo de uma pessoa do sexo masculino (feminino). Por exemplo, o nome completo Mário Henrique Simonsen seria separado da seguinte forma: Mário Henrique (nome) e Simonsen (sobrenome), pois Henrique é o primeiro termo do nome completo de milhares de meninos na nossa base. Ayton Senna da Silva, por outro lado, seria dividido como Ayrton (nome)

e Senna Silva (sobrenome), pois o termo Senna é o primeiro termo do nome completo de apenas um homem da nossa base. Elis Regina Carvalho Costa será dividido como Elis Regina (nome composto) Carvalho Costa (sobrenome) e assim por diante. Por último, na nossa análise, nomes (sobrenomes) que possuem a mesma pronúncia mas são grafados distintamente serão considerados nomes (sobrenomes) diferentes.

2..2 Saresp 2008

O banco de dados utilizado na segunda parte deste trabalho é proveniente do Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo - SARESP - realizado nos dias 27 e 28 de novembro de 2008. Essa versão do SARESP contou com a participação de 1.836.435 estudantes em 6.481 escolas. O SARESP consiste na aplicação de exames de Língua Portuguesa, Matemática, Ciências da Natureza (Biologia, Física e Química) e Redação aos alunos de segunda, quarta, sexta e oitava séries do ensino fundamental e terceira série do ensino médio das escolas públicas estaduais do estado. Além das provas aos estudantes, o Saresp 2008 aplicou questionários socioeconômicos para pais, alunos, professores e diretores. Os resultados de cada escola no Saresp 2008 foram tornados públicos e serviram para compor o Índice de Desenvolvimento da Educação do Estado de São Paulo (Idesp). Tal índice é utilizado pela Secretaria da Educação (SEE) para o diagnóstico e acompanhamento da qualidade da educação oferecida pelas escolas públicas estaduais. Além disso, o Idesp de 2008 serviu de base para a política de bonificação salarial por desempenho para professores e funcionários das escolas públicas do estado naquele ano.

As provas do Saresp 2008 foram realizadas na mesma sala e no mesmo horário em que os alunos assistiam suas aulas normais. Para aumentar a credibilidade do exame, os professores que vigiaram a execução das provas eram provenientes de outras escolas da cidade e, além disso, monitores externos foram contratados pela SEE para fiscalizar as

escolas durante a realização dos testes. O conteúdo das provas foi elaborado com base nas habilidades indicadas na Matriz de Competência e Habilidades de Língua Portuguesa, Matemática e Ciências da Natureza para cada série avaliada, a partir do Currículo elaborado pela Coordenadoria de Estudo e Normas Pedagógicas da SEE.

Os exames de Matemática e Língua Portuguesa para quarta, sexta, oitava e décima primeira séries, que serão utilizados nesta dissertação, possuíam uma parte objetiva que incluía 24 questões de múltipla escolha. Tal parte do exame foi respondida através do preenchimento de um gabarito, o qual foi corrigido através de um leitor eletrônico. Este procedimento nos garante que tais notas (notas blind) refletem única e exclusivamente a proficiência dos alunos nas matérias consideradas.

2..3 Notas do Professor

Nossa segunda fonte de dados sobre notas dos estudantes são provenientes das bases adminitrativas da SEE. Para o ano de 2008, ela compreende as notas dos boletins escolares de todos os estudantes da rede pública estadual. Tais notas estão disponíveis em período bimestral e são resultado das avaliações promovidas pelos professores dentro de sala de aula. Elas assumem valores inteiros entre 0 e 10, sendo 5 o valor mínimo que garante a progressão de série (o valor mínimo é o mesmo para todas as séries e todas as matérias). A lista de chamada, a qual deve ser preenchida diariamente pelos professores, contém o nome completo dos alunos. Portanto, os professores que atribuem tais notas têm conhecimento sobre os alunos e seus nomes completos, sendo este o motivo pelo qual chamaremos tais notas de notas non-blind. Os professores são instruídos pela SEE a avaliar os alunos segundo suas competências no currículo escolar elaborado pela Coordenadoria de Estudo e Normas Pedagógicas. Por fim, as bases administrativas contém dados sobre a frequência escolar dos estudantes.

3. Sobre os índices sociais e raciais de nomes e sobrenomes

3..1 Indicador socioeconômico

Nesta seção, apresentamos a construção do nosso índice socioeconômico de nomes e sobrenomes. Iremos caracterizar cada nome/sobrenome com base em suas frequências relativas em dois grupos distintos, são eles: alunos de escolas públicas e de escolas privadas. Ao todo contamos com o nome completo de 12.270.797 alunos. No Brasil, o ensino pré-universitário particular é um bem normal, ou seja, sua demanda cresce com a renda, refletindo o fato de que as escolas privadas são, em geral, de melhor qualidade que as públicas. A renda per capita média do aluno de escola particular, no estado de São Paulo, é cerca de 5,25 vezes a do aluno de escola pública (ver Madeira e Vicentini (2006)). A comparação das frequências dos nomes e dos sobrenomes entre alunos desses dois tipos de escola nos garante, portanto, uma boa aproximação de como tais nomes estão distribuídos entre a massa da população e a elite socioeconômica.

A vantagem desse tipo de índice é que ele nos permite distinguir de uma maneira agnóstica quais nomes/sobrenomes são de maior e menor status socioeconômico, na medida em que o aluno médio de escola privada é de maior status socioeconômico que o aluno médio do setor público. Posto isso, temos condições de avaliar quais são as características dos nomes/sobrenomes que os fazem tipicamente "ricos"(ou "pobres").

Consideremos um pai com a função utilidade $U(c, q) = c^{1-\epsilon_i}.q^{\epsilon_i}$, onde c representa consumo e q qualidade da educação do filho, e renda y_i que deve escolher entre dois tipos de escola para o filho: privada ou pública. Assuma que a qualidade da escola particular é θ_h e da pública é θ_l, tal que $\theta_h > \theta_l$. O preço da escola de melhor qualidade é p, enquanto a

pública é gratuita. Os pais são heterogêneos em duas dimensões: preferência por educação (que, provavelmente, está correlacionada com a própria educação dos pais) e renda. O parâmetro de preferência por educação é $\epsilon_i \in [a, b]$ com $a > 0$ e $b < 1$ e o de renda é $y_i \in [y_{min}, y_{max}]$, tal que $p > y_{min}$. Se $y_i \leq p$, o agente escolhe θ_l. Para $y_i > p$, o agente escolhe matricular seu filho na escola pública se, e somente se:

$$\frac{1-\epsilon_i}{\epsilon_i} \ln\left(\frac{y_i}{y_i - p}\right) \geq \ln(\theta_h) - ln(\theta_l) \tag{3.1}$$

Defina o espaço cartesiano $A = [a, b] \times [y_{min}, y_{max}]$. Os agentes que escolhem escola pública são representados pela união dos seguintes grupos: B=$\{[\epsilon_i, y_i]$, tal que $y_i \leq p\}$ e C=$\{[\epsilon_i, y_i]$, tal que $y_i > p$ e (ϵ_i, y_i) satisfaz a inequação (3.1)$\}$. O conjunto B é formado por pais cuja renda é insuficiente para pagar por θ_h. Já os pais que formam o conjunto C, por sua vez, tem renda para pagar por θ_h mas não o fazem pois não estão dispostos a sacrificar uma parcela tão grande da renda com educação de melhor qualidade. Intuitivamente, a medida que a renda cresce, a partir de p, é preciso um ϵ_i cada vez menor para os pais continuarem optando por escolas públicas. Portanto, $D = B \cup C$ formará o grupo de pais que escolhem escolas públicas e $E = A - D$ será o grupo que opta pelas particulares. O grupo D representa famílias mais pobres e/ou que valorizam menos a qualidade da educação, enquanto o grupo E é composto por famílias mais ricas e/ou com maior preferência por educação.

Consideremos agora uma função de densidade de probabilidade $f(y_i, \epsilon_i)$ definida sobre o conjunto A. Na população como um todo, a proporção de pais que escolhem o sistema público é dada por: $m = \iint\limits_D f(.)\, \mathrm{d}\epsilon\, \mathrm{d}y$. Para cada nome (sobrenome) j, defina $f_j(y_i, \epsilon_i) = f(y_i, \epsilon_i | j)$. Para o nome j, a proporção que atende ao sistema público é dada por: $m_j = \iint\limits_D f_j(.)\, \mathrm{d}\epsilon\, \mathrm{d}y$.

Com isso, definimos a fórmula do Índice Socioeconômico de Nome (Sobrenome) (ISN, ISS)

para o nome i como:

$$ISN_i = \frac{1 - m_i}{1 - m_i + \frac{(1-m)}{m} m_i} * 100 \tag{3.2}$$

Basicamente, nosso índice compara a probabilidade do aluno de nome i estar matriculado no setor privado com a probabilidade deste estar no setor público, corrigindo esta última por um fator que capta a distribuição entre pública e privada na população como um todo.

Pode-se facilmente mostrar que esse índice independe da frequência do nome na população como um todo e da fração de alunos que atendem ao sistema público. A interpretação do ISN é a seguinte: um nome terá $ISN = 100(0)$ quando todos os seus detentores estiverem no sistema privado (público). Um nome com $ISN = 50$ tem a mesma frequência nas duas populações; um com $ISN = 90$, é nove vezes mais frequente nas escolas privadas do que nas públicas; um $ISN = 10$ significa que o nome é nove vezes mais frequente nas escolas públicas do que nas privadas e assim por diante. Em suma nomes com índice de 50 serão aqueles que não apresentam conotação socioeconômica; nomes com índices menores (maiores) que 50 sao mais típicos entre crianças de menor (maior) status socioeconômico. Tal índice é o mesmo utilizado por Fryer e Levitt (2004) na análise de padrões raciais dos primeiros nomes nos EUA.

Como um exercício de robustez, para duas subamostras construímos três índices similiares ao ISN, somente mudando as categorias utilizadas nos cálculos. O primeiro deles é construído a partir da divisão da população de alunos entre beneficiários e não beneficiários do Programa Bolsa Família. O segundo e terceiro utilizam a educação do pai e da mãe, respectivamente, para dividir a população de alunos. Pais com educação acima (abaixo) da mediana formam o grupo de alta (baixa) educação. Restringindo a amostra para nomes que aparecem no mínimo vinte vezes, temos que a correlação entre os quatro índices é bastante alta, o que nos indica que o índice baseado na dicotomia escola pública/privada

capta tanto uma dimensão de renda da família do aluno quanto a dimensão de educação dos pais. A seguir temos a matriz de correlações entre os índices baseados em diferentes indicadores de status socioeconômico.

Tabela 2: Matriz de correlações dos Índices

	ISN	Educação da Mãe	Educação do Pai	PBF
ISN	1			
Educação da Mãe	.7745	1		
Educação do Pai	.7528	.9479	1	
PBF	.8493	.8331	.8138	1
N. de obs.: 1.955.868				

As correlações entre o ISN e os índices baseados no PBF e na educação dos pais são bastante altas (no mínimo .75). A seguir apresentamos também a matriz de correlações baseada em rankings construídos a partir dos índices.

Tabela 3: Matriz de correlações dos rankings dos Índices

	ISN	Educação da Mãe	Educação do Pai	PBF
ISN	1			
Educação da Mãe	.794	1		
Educação do Pai	.7752	.9554	1	
PBF	.8591	.8533	.8394	1
N. de obs.: 1.955.868				

Como a correlação entre os índices e entre os rankings dos índices são muito altas, optamos por proceder a análise usando a dicotomia pública/privada, pois nela temos a base com maior número de observações e, portanto, de nomes. Porém, cabe lembrar, que tal medida está muito correlacionada tanto com renda da família quanto com educação dos pais.

3..2 Indicador Racial

Os índices raciais de nomes e sobrenomes serão construídos a partir de uma subamostra da base usada na seção anterior. Nos exames do SARESP dos anos 2005, 2007 e 2008 os alunos tiveram que declarar nos questionários uma das seguintes raças: branco, índio, negro, mulato/pardo e asiático. Um total de 3.111.128 de alunos foram utilizados na

construção dos índices. O Índice Racial de Nome (IRN) e o Índice Racial de Sobrenome (IRS) são construídos da mesma forma que o índice socioeconômico com a diferença que as categorias agora serão branco e afro (negro e mulato). Para um nome i, o IRN_i é dado por:

$$IRN_i = \frac{P(nome_i|branco)}{P(nome_i|branco) + P(nome_i|afro)} * 100 \tag{3.3}$$

O IRS_i é construído do mesmo modo. A interpretação dos índices raciais é equivalente a dos sociais, com nomes (sobrenomes) tipicamente negros assumindo valores abaixo de 50; e tipicamente brancos, acima de 50.

4. Padrões Sociais e Raciais de Nomes e Sobrenomes

4..1 Primeiros Nomes

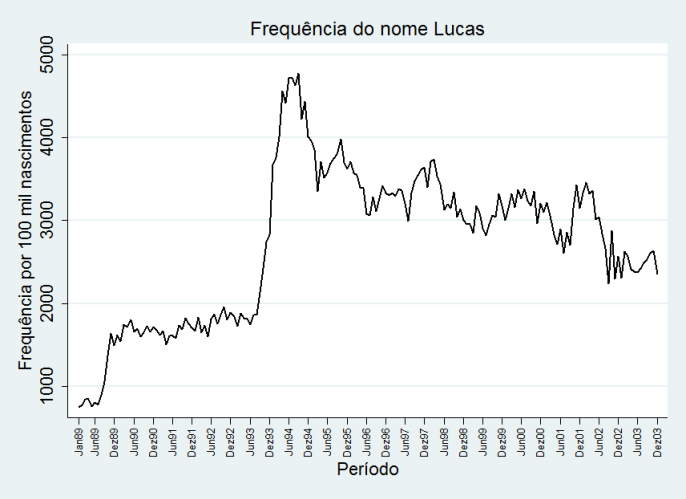
Inicialmente, a primeira caracterização que faremos dos nomes é a sua frequência estatística. Estamos interessados aqui em testar se, no Brasil, crianças de maior status socioeconômico tendem a ter nomes mais comuns. Começamos apresentando uma lista dos dez nomes mais comuns para meninos e meninas ao lado da sua frequência a cada 100 mil alunos.

Tabela 4: Nomes mais Frequentes

Meninos		Meninas	
Nome	Frequência	Nome	Frequência
Lucas	1098	Jéssica	563
Gabriel	989	Beatriz	544
Rafael	723	Leticia	521
Guilherme	691	Bruna	503
Matheus	668	Gabriela	493
Felipe	655	Amanda	469
Bruno	627	Camila	459
Vinícius	539	Larissa	445
Gustavo	534	Aline	436
Leonardo	530	Juliana	413
Frequência por 100 mil estudantes			

A lista anterior foi construída com base em alunos tanto do sistema público quanto do sistema privado. Dentre os dez nomes de meninos, cinco são de personagens bíblicos (os evangelistas Lucas e Mateus, os arcanjos Gabriel e Rafael e o apóstolo Felipe). Curiosamente, o nome mais comum entre meninos, Lucas, apresenta uma evolução temporal surpreendente. No gráfico a seguir, mostramos a evolução do número de Lucas nascidos durante os anos 90. Reparem na quebra acentuada na série que começa no período de outubro de 1993.

Figura 1: Frequência do nome Lucas no tempo



Durante o período de outubro de 1993 à junho de 1994, a frequência de Lucas nos meninos nascidos passou de 1,9 mil para mais de 4,5 mil por mil nascimentos. O que explicaria um crescimento de 137% na popularidade de um nome em tão pouco tempo? A resposta mais provável é a influência do personagem Lucas, galã da novela da 6 da Rede Globo na época. Isso é uma forte evidência de como meios de comunicação de massa ajudam a estabelecer padrões culturais entre os seus telespectadores. Para uma análise mais aprofundada, ver La Ferrara, Chong e Duryea (2008). A seguir temos as listas para os 10 nomes com maior frequência relativa para alunos de escolas públicas e particulares.

Tabela 5: Nomes de Meninos mais frequentes por tipo de Escola

Pública		Privada	
Nome	Frequência	Nome	Frequência
Lucas	1027	Gabriel	1737
Gabriel	884	Lucas	1607
Rafael	700	Guilherme	1317
Bruno	613	Matheus	1153
Felipe	607	Gustavo	1015
Guilherme	603	Felipe	996
Matheus	600	Vinícius	891
Vinícius	490	Rafael	890
Leonardo	485	Leonardo	848
Gustavo	466	Bruno	731
Frequência por 100 mil estudantes			

A análise da tabela anterior nos mostra que as listas de 10 nomes mais frequentes de meninos nas escolas públicas e privadas são compostas dos mesmos nomes. No entanto, o que salta aos olhos é a diferença das frequências relativas. Por exemplo, Lucas, o nome mais frequente para meninos estudantes do sistema público, tem uma frequência, dentro da rede pública, comparável a de Gustavo, o quinto em frequência na rede privada. Em outras palavras, a cada 100 mil alunos estudantes do sistema público, 6.475 possuem o primeiro nome entre os 10 mais frequentes; no sistema privado, este número é de 11.185. Isso é um primeiro indício de que há maior heterogeneidade de primeiros nomes entre alunos de escola pública ou, equivalentemente, de que tais alunos possuem nomes menos frequentes. Tal evidência está de acordo com a noção popular de que pais de menor renda e/ou escolaridade tendem a batizar os filhos com nomes mais incomuns. Na sequência, repetimos o mesmo exercício para os primeiros nomes de meninas.

Tabela 6: Nomes de Meninas mais frequentes por tipo de Escola

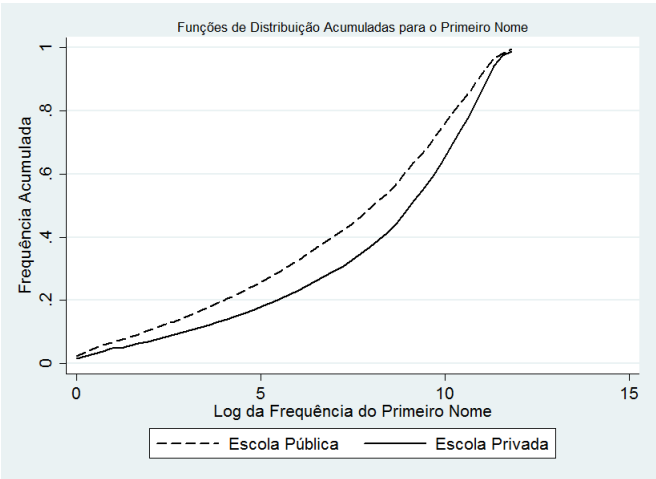
Pública		Privada	
Nome	Frequência	Nome	Frequência
Jéssica	602	Julia	1103
Bruna	488	Beatriz	1017
Beatriz	478	Gabriela	936
Leticia	477	Mariana	859
Aline	459	Leticia	830
Camila	453	Maria Eduarda	734
Amanda	453	Giovanna	691
Gabriela	431	Bruna	608
Larissa	427	Amanda	586
Juliana	403	Larissa	570
Frequência por 100 mil estudantes			

Quanto à composição das listas, temos aqui uma situação bem diferente daquela encontrada nas listas de meninos. Note que Jéssica, o nome mais comum na lista das escolas públicas, sequer aparece entre os 10 mais frequentes na escola particular; Julia, o campeão no sistema privado, também não aparece na lista das públicas. Além desses, Mariana, Maria Eduarda e Giovanna só aparecem entre os 10 na escola privada; Aline, Camila e Juliana só aparecem na lista das escolas públicas. Assim como na lista dos nomes de meninos, temos evidência de que as estudantes de escola pública tendem a apresentar

maior heterogeneidade de primeiros nomes relativamente às alunas do setor privado. A cada 100 mil alunas do setor privado, 7.934 possuem primeiro nome entre os 10 mais frequentes; esse número para a escola pública é de 4.671.

Em média, a menina estudante de escola pública tem um primeiro nome com frequência de 106 por 100 mil nascimentos; este número para meninas de escolas privadas é de 143. Para meninos, tais valores são 158 por 100 mil na pública e 234 por 100 mil na privada. A mediana para meninas (meninos) de escolas públicas é de 21 por 100 mil (29), contra 62 por 100 mil (89) no sistema privado. Por fim, um aluno de escola pública tem 11 vezes mais chance de ter um nome único (nome que só aparece uma vez na nossa base) do que um aluno do setor privado. A seguir temos um gráfico que reúne a evidência de que há maior heterogeneidade de primeiros nomes entre alunos de escolas públicas. Ele apresenta as funções de distribuição acumulada do logaritmo das frequências dos primeiros nomes para alunos de escolas públicas e alunos de escolas privadas.

Figura 2: Funções de Distribuição Acumulada para a frequência do Nome

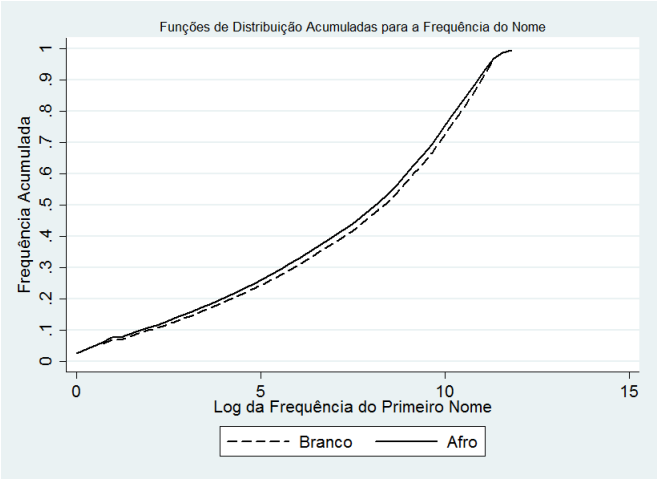


A mensagem desse gráfico é que nas escolas privadas há um universo menor de nomes e cada um deles tem maior frequência, enquanto nas públicas há um universo maior de nomes, cada qual com menor frequência. O que temos, portanto, é que pais de maior

status tendem a preferir nomes mais comuns, enquanto pais de menor status têm mais chances de adotar nomes incomuns para os filhos.

Uma pergunta natural é se a composição racial da escola pública estaria por trás de um resultado como esse. Se nomes incomuns são características exclusivas da população afrodescendente, a qual está majoritariamente matriculada no sistema público, poderíamos tomar uma diferença racial como social. Quando avaliamos só alunos de escolas públicas, praticamente não há diferença de frequência entre nomes de alunos brancos e nomes de alunos negros. A seguir, o gráfico 3 ilustra tal ponto.

Figura 3: Funções de Distribuição Acumulada para a frequência do Nome



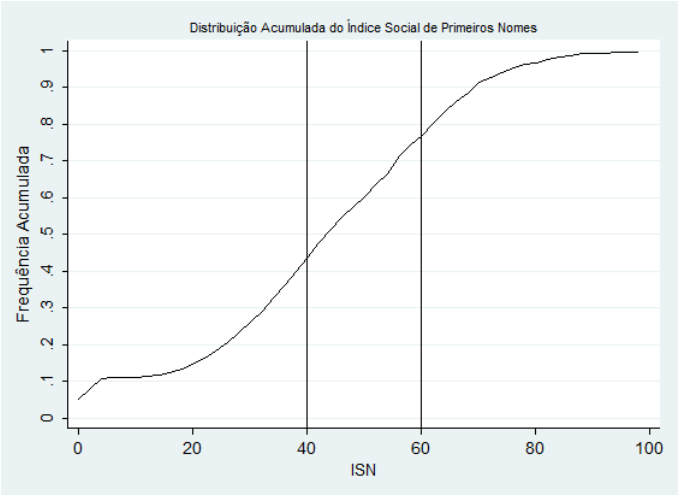
Os elementos culturais associados a escolha de nomes mais incomuns estão presentes tanto entre pais de crianças brancas quanto entre pais de crianças afrodescendentes (negras e mulatas). Isso é consistente com um cenário onde a classe social e não a raça é o elemento chave na formação do universo cultural e das relações interpessoais. Além disso, o Brasil não apresentou um movimento negro de iguais magnitudes ao norte-americano. Não deveríamos observar, portanto, a emergência de nomes tipicamente negros mas a de nomes tipicamente pobres.

São vários os fatores que contribuem para que a frequência do nome esteja correlacionada com status socioeconômico. Pais mais pobres tendem a escolher nomes mais incomuns por: uso de grafias distintas das convencionais, uso de nomes compostos, uso de nomes estrangeiros e uso de nomes inéditos.

Contudo, frequência, por si, não é uma medida completa para status pois dois nomes com a mesma frequência estatística, como Henrique e Anderson, podem ter diferentes conotações socioeconômicas. Nosso índice ISN gera uma medida de status socioeconômico para cada nome que independe da frequência do nome na população como um todo. Assim torna possível comparar status de dois nomes de frequências distintas. O ISN nos revela, de maneira agnóstica, o quão típico é um determinado nome em uma população de determinado status socioeconômico.

Se os primeiros nomes não estivessem correlacionados com status socioeconômico, a função de distribuição acumulada do ISN deveria estar próxima de uma "step function", com toda a massa da distribuição concentrada ao redor do valor 50. A seguir temos uma plotagem da função de distribuição acumulada do ISN, onde restringimos a amostra para nomes que aparecem no mínimo 20 vezes.

Figura 4: Distribuição Acumulada para o Índice Social de Nomes

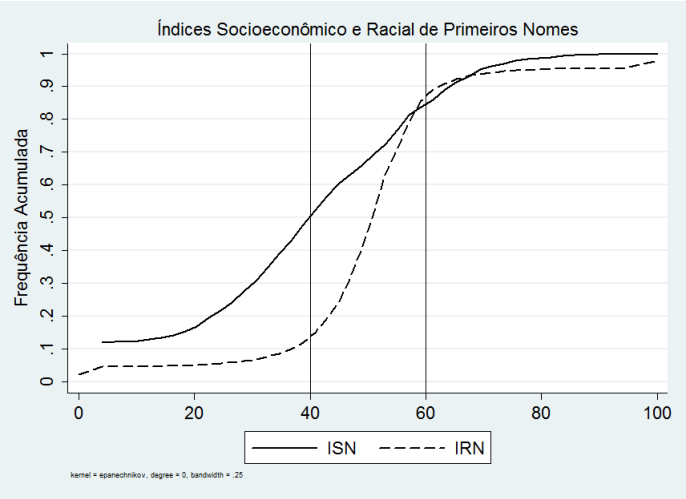


A análise do gráfico anterior revela que a distribuição do ISN está longe de ser uma step function com quebra em 50. 25% dos alunos possuem um nome que é, no mínimo, 2,44 vezes mais frequente nas escolas públicas do que nas particulares. O primeiro quartil da distribuição (percentil 75) tem ISN de 59.67. Ademais, 34% dos alunos tem um primeiro nome com índice entre 40 e 60, intervalo no qual consideramos que os nomes são praticamente neutros em termos de conotação socioeconômica. Os outros 2/3 dos alunos, pois então, possuem nomes que sinalizam algum tipo de status, alto ou baixo. Na medida em que alunos de escolas públicas e privadas caracterizam grupos de renda média distintas, portanto, podemos afirmar que o primeiro nome, de forma geral, é preditor do status socioeconômico de seu detentor.

Se primeiros nomes guardam robusta associação com status, o mesmo não é o caso entre nomes e raça. A distribuição do índice racial de nome (IRN) aponta que a esmagadora maioria dos alunos de escola pública possui um nome que está igualmente distribuído entre brancos e negros/mulatos. Em outras palavras, na nossa base o primeiro nome tem muito mais poder para predizer o status social do aluno do que para predizer sua raça. Mais uma vez a cultura dos pais revela-se mais associada à classe social que à raça pro-

priamente. O gráfico seguinte reúne as distribuições do ISN e do IRN para os alunos de escolas públicas e resume a discussão. Em especial, note os comportamentos das funções na região entre 40 e 60, onde os nomes estão igualmente distribuídos entre as categorias utilizadas.

Figura 5: Distribuições do Índice Social e do Índice Racial de Nomes



Para ilustrar a relação que geramos entre primeiros nomes e o ISN, exibimos uma sequência de tabelas com os nomes que formam os extremos da distribuição do ISN. Começamos com a lista de nomes de meninos, e seus respectivos índices, que apresentam os maiores ISN dentre os nomes que exibem mais de 1000 aparições.

Tabela 7: Nomes com maiores ISN para Meninos

Nome	ISN	Nome	ISN
Enrico	95,09	Heitor	82,86
Lucca	94,74	Ian	81,67
Luca	93,5	Thomas	80,41
Enzo	92,46	Nicholas	80,14
Bernardo	90,99	Giovanni	79,85
Luigi	89,63	Luccas	79,47
Pietro	88,50	Arthur Henrique	77,19
Pedro	85,31	Cauê	76,82
Frederico	84,25	João Pedro	76,71
Arthur	83,99	Artur	75,87

Note que entre os nomes com maiores índices há muitos de origem italiana, como Enrico,

Luca, Lucca, Enzo, Luigi, Pietro e Giovanni. O nome Enrico, por exemplo, é quase vinte vezes mais frequente nas escolas privadas que nas escolas públicas. Observe ainda que entre os vinte nomes listados apenas dois são compostos (Arthur Henrique e João Pedro). Adicionalmente, temos que uma pequena variação na grafia de Arthur está associada a uma alteração relativamente considerável do índice. Arthur é cerca de 5,25 vezes mais frequente no setor privado; enquanto Artur, cerca de 3 vezes. O nome Lucas, o mais comum entre meninos, tem índice 61,01 enquanto suas variantes que italianas têm maiores índices (Lucca 94,74, Luca 93,5). Finalmente, notamos que nenhum nome listado possui as letras K, W e Y (letras que só recentemente foram incorporadas oficialmente a Língua Portuguesa) em suas grafias.

Reduzindo o corte para o mínimo de 50 aparições, temos nomes como: Matteo (96,77), Theo (95,86), Lucas Kenzo (95,40), Kenzo (95,00), Lorenzo (94,98), Filippo (94,16), Giorgio (93,69), Theodoro (93,42), Gianlucca (92,31), Giulio (91,52), Luigi (89,63) e Ettore (87,12). Novamente, a influência italiana é bastante pronunciada - cabe lembrar que o Estado de São Paulo foi o principal destino brasileiro de imigrantes italianos entre o último quarto do século XIX e primeiro quarto do século XX. Alguns nomes típicos do Oriente Médio também têm altos índices, como: Mohamad (92,04), Ali (91,54), Tarik (88,34), Ibrahim (88,05), Mohamed (84,10) e Yasser (82,80) - provavelmente revelando a herança cultural de imigrantes sírio-libaneses, judeus e árabes. Temos ainda a influência japonesa representada pelos nomes Kenzo e Lucas Kenzo. Em suma, entre os nomes de meninos com maiores ISN, é forte a influência de nomes não-portugueses - em especial, italianos.

Compondo o outro extremo da distribuição do ISN para meninos, a seguir listamos os vinte nomes com menores índices, dentre os que apresentam mais de 1000 aparições.

Tabela 8: Nomes com menores ISN para Meninos

Nome	ISN	Nome	ISN
Elivelton	8,09	Jeferson	17,24
Romário	11,05	Alex Junior	17,5
Michael Douglas	13,17	Miquéias	18,03
Cleberson	14,60	Jackson	18,51
Cleiton	15,21	Wanderson	18,60
Gleison	15,37	Cleyton	18,82
Weverton	15,68	Adailton	19,03
Josimar	16,72	Welton	19,14
Weslei	17,07	Raimundo	19,31
Diego Aparecido	17,23	Maicon	19,51

A frequência do nome Elivelton na escola pública é cerca de 11,5 vezes maior que na escola privada. O interessante é que entre os vinte nomes listados, 12 são terminados com o sufixo ON. Além disso, 6 possuem pelo menos alguma letra estrangeira (K, W e Y). Isso revela uma tendência dos nomes tipicamente de alunos de escolas públicas de também soarem estrangeiros. Mas a influência aqui vem, predominantemente, da língua inglesa. Podemos concluir que se por uma lado os nomes que estão na cauda superior da distribuição do ISN das escolas privadas soam latinos (derivados do português ou italiano), os nomes típicos da cauda inferior soam americanizados. Outro aspecto a ser destacado é o fato do nome Maicon ser grafado de forma a reproduzir, em Português, a pronúncia inglesa do nome Michael. A tradução de Michael para o Português, Miguel tem índice de 74,48, enquanto as variantes estrangeiras tem índices piores (Maicon (19,51), Maikon (20,24), Michael (22,18), Maycon (25,46), Maickon (35,35), Michel (36,13) e Maiquel (35,35). Esse processo de empobrecimento do nome a medida que ele vai aumentando suas características estrangeiras acontece também com outros nomes como, por exemplo, Davi - Davi (61,83), David (33,88), Deyvid (31,35), Deivid (19,92), Deividi (12,89) e Deivide (9,76).

Reduzindo o corte para no mínimo 50 aparições, teremos nomes como: Welisson (3,51), Geilson (3,99), Joedson (4,05), Wallas (4,22), Claudiano (4,33), Nadson (4,52), Wesley Bruno (4,61), Jefferson Lucas (4,83), Mailon (5,03) e Maicon Cesar (5,1). Todos são, no mínimo, 19 vezes mais frequentes nas escolas públicas que nas escolas particulares. Novamente, os nomes apresentam forte influência da língua inglesa, representado pela presença

de W e Y, além do sufixo son (presente em sobrenomes de origem anglo-saxã com o significado "filho de"). Há, portanto, evidência de que a cultura de massa norte-americana é especialmente atrante entre os brasileiros mais pobres quando da escolha do primeiro nome.

A história do nome Romário, o segundo nome da tabela acima, revela a influência até mesmo do futebol sobre as escolhas dos pais. A maioria dos Romários da nossa base nasceu justamente em junho de 1994, período em que o atacante de mesmo nome se destacou na conquista da Copa do Mundo de futebol. Como Romário hoje é um nome de baixo status, concluímos que pais mais pobres e/ou menos educados estão mais propensos a realizar esses tipos de homenagens.

Para meninas, a lista dos vinte nomes com maiores índices, dentre os que aparecem mais de mil vezes, é a seguinte:

Tabela 9: Nomes com maiores ISN para Meninas

Nome	ISN	Nome	ISN
Luisa	92,89	Ana Luisa	85,42
Sofia	92,00	Catarina	85,23
Pietra	90,05	Laura	83,97
Sophia	89,99	Lara	83,3
Giulia	89,55	Isabella	82,63
Clara	89,11	Maria Luisa	82,33
Luiza	88,32	Marina	82,30
Manuela	87,70	Marcella	82,29
Manuella	86,93	Helena	82,10
Giuliana	85,88	Julia	82,00

O primeiro nome da lista, Luisa, é quase 13 vezes mais frequente nas escolas particulares do que nas públicas; este valor é de 4,5 para Julia, o nome que fecha a lista. Em comparação com a lista dos nomes masculinos, a influência estrangeira é menos acentuada (ficando por conta dos italianos Pietra, Isabella (correspondente de Isabel), Giulia (italiano de Julia) e Giuliana (italiano de Juliana)). Dos 20 nomes listados, apenas 2 são compostos - Ana Luisa e Maria Luisa. Fazendo o mesmo exercício, mas reduzindo o corte para o

mínimo de 50 aparições temos os seguintes nomes: Nina (96,78), Valentina (95,68), Maya (95,08), Chiara (92,89), Mariah (92,15), Sophie (91,25), Lia (90,74), Giorgia (90,55), Liz (90,74) e Maithe (89,79). Todos são, no mínimo, 10 vezes mais frequentes nas escolas particulares que nas públicas.

Por outro lado, a seguir temos a lista dos nomes de meninas com menores índices, dentre aqueles com o mínimo de 1000 aparições.

Tabela 10: Nomes com menores ISN para Meninas

Nome	ISN	Nome	ISN
Carolaine	10.59	Géssica	18.00
Tamires Aparecida	12.54	Andressa Aparecida	18.09
Karolaine	14.48	Daiane	18.91
Josefa	14.52	Marinalva	19.34
Graciele	14.61	Leidiane	19.39
Jéssica Aparecida	14.62	Raimunda	20.11
Gleice	17.14	Adriele	20.22
Crislaine	17.17	Maria Socorro	20.50
Daiane Aparecida	17.19	Eliene	20.76
Franciele	17.88	Pamela Cristina	21.20

Carolaine, o primeiro da lista, é cerca de 9 vezes mais comum nas escolas públicas que nas particulares. Assim como o caso de Maicon, Carolaine é grafado de modo a reproduzir no português a pronúncia em inglês de Caroline. O mesmo ocorre com Daiane (reproduz a pronúncia de Diane) e, possivelmente, Leidiane (Lady Diane). A lista termina com Pamela Cristina (5 vezes mais provável nas públicas). Note ainda que dos 20 nomes, 6 são nomes compostos, sendo 4 terminados em Aparecida (nome da santa padroeira do Brasil na religião católica, a majoritária no país). Repetindo o exercício mas reduzindo o corte para no mínimo 50 aparições temos: Lenilda (4,27), Crislane (4,34), Santa (4,70), Juciele (4,76), Maria Daiane (5,07), Jéssica Daiane (5,07), Géssica Cristina (5,38), Natasha Aparecida (5,42), Janeide (5,46) e Benedita Maria (5,46). Tais nomes são, no mínimo, 15,66 vezes mais frequentes nas escolas públicas que nas escolas privadas. Destaque para Géssica, uma variante gráfica de Jéssica - o nome mais comum entre meninas.

Modelamos o ISN e o IRN usando como variáveis independentes indicadores de educação dos pais, raça da criança, do pai, da mãe e os índices raciais e sociais dos sobrenomes herdados do pai e da mãe. Com isso podemos testar para a existência de relação entre educação dos pais e status dos primeiros nomes e saber se o sobrenome tem influência na escolha do nome. A seguir os resultados:

Tabela 11: Determinantes do ISN e do IRN

Variável	ISN	IRN
Mãe c/ e. superior	3.4	.86
Pai c/ e. superior	2.58	.68
Mãe c/ e. médio	.3	.3
Pai c/ e.médio	.39	.14
Mãe s/ escola	-.29	-.36
Pai s/ escola	-.82	-.66
Criança Afro	-.92	-8.53
Pai Afro	.046	-.76
Mãe Afro	.046	-.98
$ISS_{mãe}$	.046	.003
ISS_{pai}	.046	.002*
$IRS_{mãe}$	.046	.003
IRS_{pai}	.037	.033
Meninas	-2.48	3.18
Ano de Nascimento	.89	.36
Obs.	980,971	980,971

Todos os coeficientes são significantes a 1 %, exceto *, não significante

A tabela anterior nos revela que as variáveis independentes usadas estão mais associados ao ISN do que ao IRN. Tal ponto já era esperado pois a quantidade de variação do ISN a ser explicada é bem superior a do IRN. Meninos com pais com diploma universitário têm em média nomes mais "ricos", ao passo que meninas brancas têm em média nomes mais "brancos". Os índices de sobrenome tem um impacto tímido sobre os índices de nome. Uma variação de 1 ponto em cada um dos quatro índices de sobrenome leva a um aumento de .175 do ISN, o qual está numa escala de 0 a 100. A relação entre os índices de sobrenome e o IRN é ainda mais fraca.

Uma vez controlado por educação dos pais e sobrenomes, praticamente não há diferença entre o status social dos primeiros nomes de crianças brancas e afrodescendentes. A diferença do ISN é da ordem de .92, menos de um ponto. Por construção, crianças afro tem

primeiros nomes também mais "afro", sendo o efeito da ordem de 8,53 pontos. Novamente, os dados indicam que nomes de pobre são características culturais de pais pobres, sejam eles brancos ou afrodescendentes.

Dado que em se tratando de primeiros nomes no Brasil, a dimensão relevante é o status social e não a raça, passamos para o estudo das características dos nomes que os tornam tipicamente pobres. Há evidência anedótica de que no Brasil o uso de letras estrangeiras (K, W e Y) tanto no começo dos nomes quanto em substituição às letras C, U e I no meio do nomes empobrece os primeiros nomes. O mesmo aconteceria com a substituição do W pelo U em nomes como William, ou do K pelo Q em nomes como Kelvin. Além disso, o uso do sufixo son no final de nomes masculinos também indicaria nomes pobres. A origem de tal sufixo remonta aos sobrenomes patronímicos originários da língua inglesa, como Johnson (filho de John), Jackson (filho de Jack), entre outros. Testamos tais conjecturas e, a seguir, dispomos uma tabela com os coeficientes das regressões do ISN contra características dos primeiros nomes. As regressões são feitas separadamente para meninos e meninas.

Tabela 12: Regressões do ISN contra características ortográficas dos Nomes

	Meninas		Meninos	
	dummy	constante	dummy	constante
K inicial	-10,79	42,10	-2,24	43,37
K meio	-18,28	42,03	-13,26	43,58
W inicial	-19,75	41,78	-20,52	44,17
W meio	-19,92	41,76	-21,17	43,58
Y inicial	10,11	41,67	8,8	43,28
Y meio	-9,90	42,78	-16,51	43,96
Q inicial	-20,90	41,76	-27,93	43,33
Q meio	-10,59	41,92	2,33	43,15
U inicial	-17,64	41,74	-21,77	43,36
U meio	3,35	41,17	8,07	40,13
Sufixo ON			-21,83	44,74
Sufixo SON			-21,15	44,2
N. Obs	6.121.920		6.148.877	
Todos os coeficientes são significantes a 1%				

Em média, um nome de menino ou menina começado em W tem um ISN 48% menor do que aqueles com outras letras iniciais. Um W no meio do nome, seja para substituir o U

ou V, também empobrece o nome na mesma medida. Enquanto nomes começados com Y têm maiores ISN, o contrário vale quando o Y é encontrado no meio do nome substituindo a letra I. Padrão oposto vale para a letra U, sugerindo que a inicial U está sendo usada para reproduzir o som da inicial W. São os casos de Ualace e Ueslei, por exemplo. A presença das letras K e Q também está associada a baixo status, especialmente quando são iniciais. Um nome de menino com inicial Q, por exemplo, tem ISN 64% menor.

Nomes de meninos terminado em SON ou ON tem o ISN reduzido, em média, em cerca de 48%. Por que os nomes com tais estrangeirismos tem apego tão grande entre a população brasileira mais pobre? Uma conjectura plausível é que a cultura de massa (cinema, televisão e música), especialmente a produzida em língua inglesa, tem servido de modelo de sucesso para pais brasileiros mais pobres e/ou menos educados. Estudos de caso como a popularização de nomes como os da princesa Diana ou Lady Di (com grafias como: Daiana, Leidiana) e os de atores de Hollywood como Michael Douglas confirmam que a cultura pop é especialmente atraente nas camadas mais pobres da população brasileira. O fenômeno é tão pervasivo no Brasil que motivou o deputado federal Paulo Magalhães (Democratas, Bahia) a redigir o Projeto de Lei 7520/10, o qual pretende alterar a Lei de Registros Públicos (lei número 6015 de 31 de Dezembro de 1973), de forma a proibir os cartórios de registrar crianças com nomes estrangeiros. A seguir dispomos os vinte nomes mais comuns terminados no sufixo SON e seus respectivos ISN. Notem que o maior ISN da lista é ainda assim menor que 50.

Tabela 13: Nomes mais frequentes terminados no sufixo SON

Nome	ISN	Nome	ISN
Anderson	27.94	Gilson	30.90
Jefferson	21.47	Denilson	20.30
Robson	25.39	Wanderson	18.60
Emerson	25.66	Edilson	26.54
Jeferson	17.24	Nelson	45.53
Edson	34.12	Alison	28.72
Alisson	26.04	Edmilson	29.88
Wilson	38.59	Gerson	34.46
Jackson	18.51	Peterson	32.27
Adilson	33.32	Jailson	13.39

Os dados indicam também que nomes compostos, em geral, estão associados a baixo status. Um nome masculino composto tem, em média, um ISN 21% menor. A título de exemplo, dois nomes muito comuns como Thiago (53,86) e Rodrigo (52,06), nomes praticamente neutros em conotação socioeconômica, quando compostos empobrecem (o nome composto Thiago Rodrigo tem ISN de 32,15). Para meninas, nomes compostos tem ISN cerca de 30% menor. Como exemplo, temos os nomes Bruna (55,44) e Camila (52,25) que, quando unidos, geram o nome composto Bruna Camila, que tem ISN de 34,32. Entre os nomes compostos, destaque para aqueles em que os nomes Aparecida ou Aparecido aparecem como segundo termo. Tais nomes têm seus ISNs reduzidos em 22% para meninas e 42% para meninos, revelando que esse tipo de costume religioso é mais típico entre as classes mais baixas.

4..2 Sobrenomes ou Nomes de Família

Agora começamos nossa análise de sobrenomes. O estudo de sobrenomes é recorrente em áreas como genealogia e genética populacional. Aqui, vamos utilizar os dados para fazer, assim como foi feito com os nomes, uma descrição dos padrões socioeconômicos e raciais dos nomes de família.

Enquanto nomes refletem uma escolha dos pais, o mesmo não vale para os sobrenomes. Os agentes podem escolher seu cônjuges, mas, uma vez definidos os pareamentos no mercado

de casamentos, os nomes de família viram estoque e não variáveis de escolha. Junto a isso, os sobrenomes são passados de geração em geração - no Brasil, em geral o filho herda os sobrenomes dos avós paterno e materno. Tais fatores indicam que sobrenomes captam características familiares que se perpetuam no tempo, ao contrário dos primeiros nomes, os quais estão sujeitos a modas e choques culturais de curto prazo. O status socieconômico de um sobrenome, portanto, pode ser entendido como uma medida do sucesso de longo prazo da família que o carrega, enquanto o status racial revela a posição familiar na formação histórica da sociedade brasileira.

Dois eventos históricos fundamentais contribuíram para a formação do universo de sobrenomes que observamos hoje na sociedade paulista. O primeiro deles é a abolição do trabalho escravo em 1888. Até então, o que se tinha era a sociedade branca que carregava sobrenomes portugueses e sobrenomes derivados dos apelidos índigenas e negros das propriedades rurais, estes últimos adotados por famílias ilustres no período nacionalista que seguiu à Independência brasileira (1822). Parte dos negros adotaram os sobrenomes de seus senhores, com o consentimento desses. A partir de 1888, os ex-escravos precisavam de algum sobrenome para realizar seu registro civil. Muitos escolheram sobrenomes religiosos católicos. Outros escolheram sobrenomes tipicamente portugueses, como uma forma de reafirmar sua igualdade perantes os brancos.

O segundo evento histórico está intimamente relacionado ao primeiro. Com a abolição da escravatura, a mão-de-obra negra passou a dar espaço para os trabalhadores imigrantes de diversos países. O Estado de São Paulo foi o principal atrator brasileiro de levas de imigrantes italianos, japoneses, portugueses, espanhóis, sírio-libaneses e de diversas outras nacionalidades. Segundo o Atlas da Imigração Internacional em São Paulo, entre 1890 e 1900, a parcela de estrangeiros na população paulista passou de 5,4% para 21%. Nesses dez anos, o número de imigrantes vivendo em São Paulo passou de 75.030 para 478.417. São Paulo continuou atraindo imigrantes, embora em menor escala, durante a primeira

metade do século XX.

Enquanto o registro de negros ex-escravos contribuiu para a massificação dos sobrenomes portugueses, a presença dos imigrantes introduziu uma grande variedade de sobrenomes na sociedade paulista. É certo, portanto, que os padrões sociais e raciais dos sobrenomes do Brasil contemporâneo estão fortemente conectados a tais ocorrências históricas.

Iniciamos nossa análise avaliando a frequência estatística dos sobrenomes na nossa base. Nosso primeiro objetivo é avaliar, assim como fizemos no caso dos primeiros nomes, se existe associação entre o status social e racial dos sobrenomes e suas frequências estatísticas. A seguir dispomos uma tabela com os 20 sobrenomes mais comuns, ao lado da porcentagem dos alunos que possuem tal sobrenome.

Tabela 14: Sobrenomes mais Frequentes

Sobrenome	Frequência (%)	Sobrenome	Frequência (%)
Silva	20,80	Gomes	2,10
Santos	13,89	Almeida	2,01
Oliveira	7,86	Nascimento	1,99
Souza	6,14	Ribeiro	1,88
Pereira	4,08	Araujo	1,69
Ferreira	3,61	Sousa	1,57
Rodrigues	3,50	Barbosa	1,55
Alves	3,45	Gonçalves	1,54
Lima	3,38	Martins	1,51
Costa	2,21	Carvalho	1,44

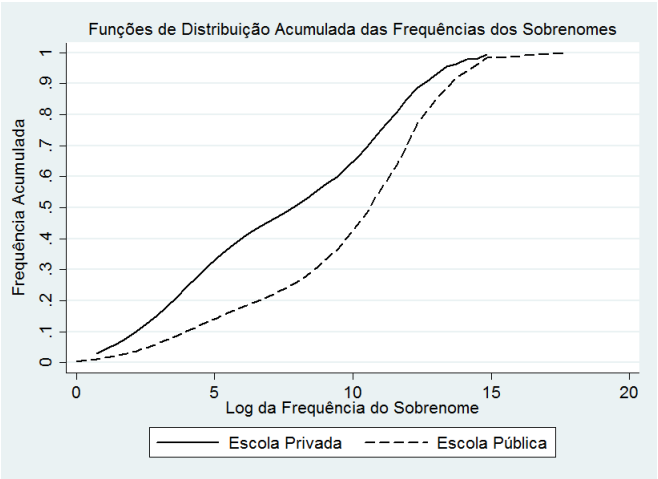
Refletindo a influência majoritariamente lusitana, todos os 20 sobrenomes listados anteriormente são portugueses. O impressionante da tabela anterior é a prevalência dos sobrenomes Silva e Santos. A cada cinco alunos, um tem Silva no sobrenome. A cada dez alunos, pelo menos um tem Santos como nome de família. Para efeitos de comparação, na Espanha, o sobrenome mais frequente é Garcia, com 3,5% da população. Tal frequência é equivalente a de Rodrigues, o sétimo na nossa lista. Na Inglaterra e País de Gales, o sobrenome Smith é o mais comum, com 1,5%. Isso é comparável ao sobrenome Martins, o penúltimo na lista. Como dito anteriormente, nossa base está muito próxima do nível

censitário, logo tais resultados valem para toda a população paulista nascida durante os anos 90.

A tabela anterior suscita duas questões. A primeira seria o motivo pelo qual os sobrenomes Silva e Santos estão tão descolados dos demais em termos de frequência. A segunda seria o porquê o Brasil apresenta uma concentração tão grande de indivíduos num universo tão pequeno de sobrenomes quando comparado a outros países. Uma cenário plausível para a explicação desses padrões é um em que exista adoção de sobrenomes. A adoção dos sobrenomes dos senhores por parte dos ex-escravos explicaria a grande popularidade dos sobrenomes portugueses em geral, enquanto a adoção em massa pós-1888 explicaria a enorme concentração relativa em Silva e Santos.

Uma pergunta natural seria a de se a frequência estatística de um sobrenome dá algum sinal nos dias de hoje sobre o status social de um indivíduo. Investigamos tal pergunta avaliando a distribuição de frequências dos sobrenomes entre alunos do sistema público e particular de ensino. A seguir temos um gráfico com a plotagem dessas distribuições.

Figura 6: Funções de Distribuição Acumulada das Frequências dos Sobrenomes

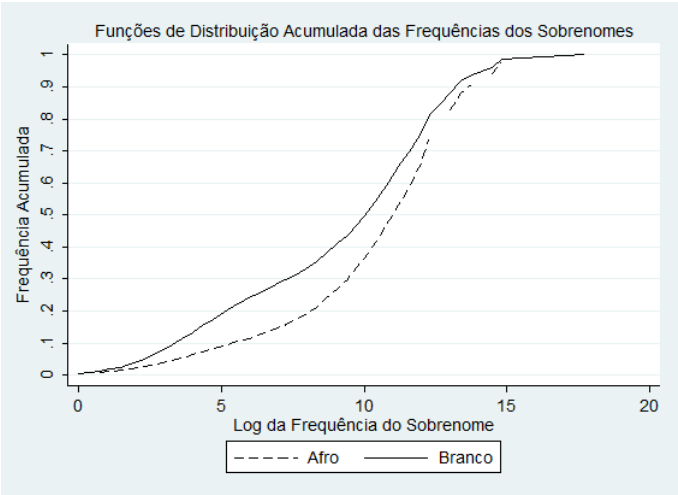


O gráfico acima revela um padrão contrário ao observado no caso dos nomes. Enquanto ter

um primeiro nome mais incomum está associado a baixo status, ter um sobrenome mais incomum está associado a maior status socioeconômico. Reparem que cerca de 70% dos alunos do sistema privado têm um sobrenome mais incomum que o sobrenome do aluno mediano da escola pública. Tal resultado é consistente com a existência de inércia na transmissão intergeracional de status econômico. Em outras palavras, se no passado ter um sobrenome comum estava associado a condições sociais menos favoráveis, o mesmo se passa no presente. Nesse ponto, vale a pena ressaltar as diferenças entre o caso brasileiro e o espanhol. Enquanto na Espanha a relação entre frequência do sobrenome e status se estabeleceu por conta do mecanismo de sinalização (famílias de sucesso hifenizando os sobrenomes), no caso brasileiro a história escravocrata e a imigração em massa são as forças primordiais.

Quando olhamos só para alunos do setor público e comparamos as frequências dos sobrenomes entre brancos e afrodescendentes, obtemos também uma situação distinta da dos primeiros nomes. Enquanto o padrão de frequência dos nomes é similar entre alunos das duas raças, alunos brancos tendem a ter sobrenomes mais incomuns que alunos afrodescendentes, como evidenciado pelo gráfico a seguir.

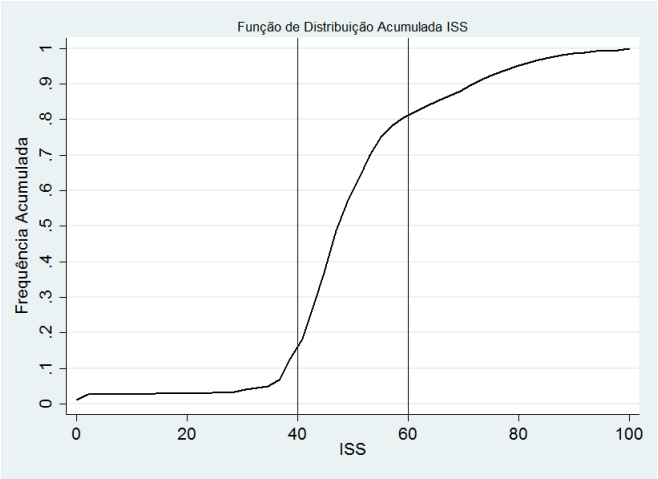
Figura 7: Funções de Distribuição Acumulada das Frequências dos Sobrenomes



Em resumo, os dados até agora retratam o seguinte quadro: alunos do setor privado possuem menor variedade de nomes, nomes mais comuns, maior variedade de sobrenomes e sobrenomes mais incomuns; alunos de escolas públicas apresentam maior variedade de nomes, nomes menos comuns, menor variedade de sobrenomes e sobrenomes mais comuns. Em termos de raça, os sobrenomes mais incomuns são mais típicos entre os alunos brancos. Dentro da escola pública, não há diferença na frequência dos nomes entre alunos brancos e afrodescendentes.

O Índice Social do Sobrenome (ISS) e o Índice Racial de Sobrenome (IRS) apresentam alta correlação. Isso quer dizer que, em geral, nomes tipiciamente ricos são tipicamente brancos e nomes tipicamente pobres são tipicamente negros. A seguir, apresentamos uma plotagem da distribuição do ISS entre os alunos de todo o sistema de ensino pré-universitário de São Paulo. Como a maioria dos alunos possui mais de um sobrenome, o ISS de cada aluno foi construído como uma média ponderada dos ISSs dos sobrenomes individuais do aluno, onde os pesos de cada sobrenome são determinados pelo inverso de sua frequência estatística. Portanto, quanto mais raro for um sobrenome com relação aos outros que formam o sobrenome completo do indivíduo, maior será o peso de seu ISS individual na formação do ISS do índivíduo. O ISS do sobrenome Senna da Silva, por exemplo, será uma média ponderada entre os ISSs dos sobrenomes Senna e Silva, onde Senna, por ser muito mais raro que Silva, terá maior peso.

Figura 8: Função de Distribuição Acumulada do Índice Social de Sobrenome



A distribuição para o IRS segue um padrão parecido. Segundo o gráfico, cerca de 65% da população tem um sobrenome completo que não possui forte conotação social. São os sobrenomes completos que apresentam ISS entre 40 e 60. Cabe lembrar que nesse cálculo, nomes de massa como Silva e Santos - tipicamente pobres - tem peso reduzido justamente por serem extremamente comuns. Por outro lado, 20% da população possui sobrenome completo que indica maior status, enquanto 15% tem um sobrenome completo tipicamente de baixo status.

A partir dessas informações, a questão que se põe é se existe algum padrão entre os sobrenomes que ocupam o piso, o meio e o topo da distribuição do ISS. Para isso, o primeiro passo é identificar a provável origem do sobrenome. Faremos tal identificação através da análise da grafia de cada sobrenome: sobrenomes cujas grafias diferem da Língua Portuguesa formal serão considerados estrangeiros. Nosso algoritmo de classificação de sobrenomes começa com a classificação manual dos 12 mil sobrenomes mais frequentes. Estes são classificados em portugueses e outros. Para os sobrenomes classificados como outros, avaliamos a presença de sufixos que dão uma pista sobre a provável origem do sobrenome (Itália, Sírio/Libano, Japão, Espanha, Europa não-latina e Portugal). Por exemplo, sobrenomes terminados em sufixos como Gia, Agi, Atti, Ecco, Azza, Ini, Ucci,

One, Onne, Ato, Etto, Olo serão classificados como de origem italiana (são os casos de Peloggia, Zambiagi, Menegatti, Grecco, Corazza, Scottini, Bertolucci, Donadone, Spoletto, Fasolo, entre outros). Sobrenomes terminados com Ssar, Ssur, Ad, Ef, Af, Ab, Eb serão classificado como sírio-libaneses (como exemplo, temos Nasser, Farah, Maluf, Afif e Abud). Adotamos o mesmo procedimento para classificar sobrenomes como japoneses (sufixo Moto, vide Yamamoto, sufixo Gawa, vide Nakagawa) e de outras partes do continente europeu (er, como Muller; mann, como Zimmermann).

Ao todo temos 306.885 sobrenomes diferentes, onde 17.400 são identificados como portugueses, 47.451 italianos, 2.697 sírio-libaneses, 1.047 espanhóis, 4.400 japoneses, 50.437 da europa não-latina e 183.453 que ficaram sem identificação. Se uma pessoa da nossa base tem n sobrenomes, ela passa a contar como n "pessoas virtuais", cada uma dessas representando um dos sobrenomes. O nome Edson Arantes Nascimento viraria duas pessoas: uma com sobrenome Arantes e outra com sobrenome Nascimento. Com isso, temos que a contagem total de sobrenomes na nossa base é de 23.225.858, sendo 20.481.678 portugueses, 834.256 italianos, 34.992 espanhóis, 383.672 do resto da europa, 24.039 sírio-libaneses, 77.811 japoneses e 1.389.410 não identificados.

A seguir apresentamos uma tabela com a classificação nacional do sobrenome, seu ISS médio, sua parcela correspondente no conjunto de sobrenomes e sua parcela na população.

Tabela 15: Média do ISS por origem

Origem	Média ISS	% dos Sobrenomes	% da População
Italiana	67,51	15,46	3,59
Sírio-Libanesa	74,01	0,87	0,1
Espanhola	70,17	0,34	0,15
Portuguesa	45,28	5,66	88,18
Japonesa	78,82	1,43	0,3
Resto da Europa	60,3	16,43	1,70
Não identificado	66,80	59,81	5,98

A análise da tabela anterior nos revela que sobrenomes portugueses estão aproximada-

mente igualmente distribuídos entre alunos de escolas públicas e particulares. Em outras palavras, sobrenomes portugueses, em média, não dão sinal sobre origem socioeconômica. A situação é diferente, entretanto, para sobrenomes de origem não-portuguesa. Em média, um sobrenome japonês tem um índice 74% maior que um sobrenome português; um sobrenome sírio-libanês, 64%; um sobrenome espanhol ou italiano, 50%. O padrão apresentado pelos dados é de que os sobrenomes portugueses são, em média, mais pobres e mais negros que os sobrenomes de descendentes de imigrantes. A título de ilustração, selecionamos alguns sobrenomes patronímicos (sobrenomes derivados do nome do pai, como Fernandes, que significa filho de Fernando) portugueses e espanhóis. Os patronímicos portugueses são terminados em Es, enquanto os espanhóis em Ez. A diferença de uma letra está relacionada a grandes diferenças nos índices social e racial.

Tabela 16: Patronímicos Portugueses e Espanhóis

Português	ISS	IRS	Espanhol	ISS	IRS
Lopes	50,29	50,49	Lopez	79,90	66,14
Rodrigues	45,87	49,28	Rodriguez	80,70	68,33
Fernandes	54,00	55,67	Fernandez	79,05	68,85
Soares	44,48	46,53	Suarez	83,83	72,10
Gomes	44,13	46,16	Gomez	73,04	67,36
Vasques	68,06	69,76	Vasquez	74,55	80,36
Gonçalves	48,21	49,33	Gonzalez	79,84	72,88
Sanches	67,72	72,00	Sanchez	75,61	74,67
Alvares	66,83	69,96	Alvarez	79,66	71,70
Marques	53,73	52,21	Marquez	72,62	55,88

Enquanto os sobrenomes portugueses estão democraticamente distribuídos entre pobres e ricos e entre brancos e afrodescendentes, os de origem espanhola são tipicamente de maior status e tipicamente brancos. Esse quadro se mantém quando olhamos para sobrenomes de outras origens. A seguir apresentamos o ISS para os dez sobrenomes mais frequentes por origem imigratória. Os dados sugerem que, em São Paulo, sobrenomes de origem não-portuguesa estão fortemente associados a maior staus social.

Tabela 17: Sobrenomes mais frequentes por origem

Italiano	ISS	Resto da Europa	ISS	Sírio-Libanês	ISS	Japonês	ISS
Rossi	73.67	Bittencourt	71.11	Haddad	91.70	Hideki	86.61
Ferrari	73.19	Furlan	75.74	Dib	88.11	Tanaka	84.06
Mantovani	70.38	Trevisan	72.04	Farah	88.85	Yamamoto	83.90
Netto	72.73	Muller	70.8	Mansur	87.5	Suzuki	84.20
Bianchi	72.96	Schmidt	78.33	Saad	92.51	Nakamura	85.84
Sartori	73.74	Fisher	73.42	Maluf	91.54	Takahashi	83.47
Moretti	73.85	Bressan	72.59	Said	78.52	Hiroshi	77.22
Milani	69.79	Keller	61.14	Nasser	91.53	Yamada	83.14
Marchi	72.83	Schneider	75.12	Abud	89.22	Yukari	86.66
Ricci	71.27	Stein	74.40	Nassif	87.32	Yoshida	85.72

Outra forma de avaliar se sobrenomes estrangeiros estão associados a maior status socioeconômico é estudar a correlação entre o ISS e a presença de letras estrangeiras (K, W e Y) na grafia do sobrenome. Para isso, rodamos regressões do ISN contra dummies de presença de pelo menos uma letra estrangeira na grafia e obtemos os seguintes resultados:

Tabela 18: Letras Estrangeiras no Sobrenome e ISS

	Constante	Dummy
Pelo menos um K	47,08	28,78
Pelo menos um W	47,19	24,45
Pelo menos um Y	47,13	20,50

Ao contrário do que ocorre com os nomes, ter uma letra estrangeira no sobrenome está correlacionado com maiores ISS. Ter pelo menos um K ou um W no sobrenome aumenta o índice socioeconômico do sobrenome em mais de 50%. Para a letra Y esse valor é de cerca de 43,5%. Realizando o mesmo exercício para as outras letras do alfabeto, concluímos que a presença de cada uma delas na grafia do nome não está correlacionada com mudanças significativas no ISS. Em resumo, enquanto a presença de letras estrangeiras nos primeiros nomes está correlacionada negativamente ao ISS, nos sobrenomes tal presença está associada a maiores ISS. Enquanto pais de baixo status socioeconômico tendem a escolher nomes com letras estrangeiras para seus filhos, pais de alto status passam para seus filhos sobrenomes com escritas estrangeiras.

Os sobrenomes portugueses mais comuns da nossa base não estão associados a status so-

cial e racial, com a importante exceção dos sobrenomes de massa Silva e Santos. Como podemos ver na tabela a seguir, dentre os mais comuns apenas três estão fora do intervalo 40-60, dentro do qual podemos considerar o nome neutro de conotação social e racial. Enquanto os devocionais Santos e Nascimento são tipicamente pobres e negros, Silva só é tipicamente pobre.

Tabela 19: Sobrenomes Portugueses mais Frequentes e seus Índices

Sobrenome	ISS	IRS	Sobrenome	ISS	IRS
Silva	36.24	45.59	Gomes	44.13	46.16
Santos	35.83	37.63	Almeida	47.61	49.33
Oliveira	44.23	47.33	Nascimento	38.77	39.31
Souza	42.35	43.42	Ribeiro	48.41	48.71
Pereira	41.25	43.67	Araújo	44.07	46.66
Ferreira	43.80	45.74	Sousa	35.16	42.79
Rodrigues	45.87	49.3	Barbosa	43.05	45.52
Alves	41.23	46.07	Gonçalves	48.21	49.33
Lima	42.46	47.03	Martins	52.60	52.54
Costa	47.07	46.90	Carvalho	53.22	51.00

Os sobrenomes Silva, Santos, Sousa e Nascimento, por exemplo, são 80% mais comuns entre alunos de escolas públicas do que entre estudantes de instiuições particulares, enquanto Santos e Nascimento são cerca de 60% mais frequentes entre alunos afrodescendentes. Os outros nomes de família da lista não apresentam tendência clara sobre o tipo da instituição do estudante nem sobre sua raça. De forma geral, o próprio processo de adoção dos sobrenomes portugueses pelos brasileiros fez com que tais nomes de família perdessem sua capacidade de sinalizar posição social e raça. Existem, contudo, alguns sobrenomes de origem lusa com maiores índices. Tais nomes tendem a ser menos frequentes e apresentam padrões semelhantes aos sobrenomes estrangeiros. Alguns deles, dentre os que possuem um mínimo de mil aparições, são: Junqueira (81,46), Meirelles (76,48), Valente (75,52), Henriques (73,91) e Vasconcelos (73,81).

Identificados os tipos de sobrenome que ocupam o topo e o meio da distribuição do ISS e IRS, resta-nos investigar os tipos de nomes de família que ocupam a cauda inferior. Como dito, trabalhos em genética no Brasil demonstram correlação entre traços genéticos típicos

de populações africanas e sobrenomes devocionais católicos. Tal correlação se estabeleceu pela necessidade de registro civil de negros libertos, os quais podiam escolher o sobrenome. Dado que o universo cultural destes era fortemente influenciado pelo catolicismo, muitos negros optaram por sobrenomes com menção religiosa. O ponto interessante da nossa investigação é avaliar se ainda hoje, várias gerações depois da adoção, tais sobrenomes continuam pobres e negros. Já vimos anteriormente o caso do sobrenome Santos. A seguir listamos vinte sobrenomes derivados do catolicismo e seus respectivos índices.

Tabela 20: Sobrenomes Devocionais e seus Índices

Sobrenome	ISS	IRS	Sobrenome	ISS	IRS
Dores (das Dores)	27.59	41.98	Adão	37.31	30.50
Graças (das Graças)	28.23	42.20	Batista	38.69	44.7
Virgens	28.38	34.6	Nascimento	38.77	39.31
Conceição	28.87	31.83	Santana	38.83	37.92
Aparecido	28.89	47.25	Anjos (dos Anjos)	39.08	39.75
Remédios (dos Remédios)	29.57	23.67	Evangelista	41.4	42.45
Jesus (de Jesus)	29.79	31.78	Cruz	41.53	40.48
Aparecida	30.39	49.00	Deus	41.69	42.47
Anunciação	33.24	28.39	Paixão	41.79	34.24
Bispo	33.88	30.29	Messias	42.81	46.00

Notemos que os 7 sobrenomes com menores ISS são cerca 2,5 vezes mais frequentes nos colégios públicos do que nos particulares. Em outras palavras, são cerca de 150% mais prováveis em instituições públicas. Os sobrenomes da lista anterior também estão mais concentrados nas populações afrodescendentes. O nome de família Jesus ou de Jesus, por exemplo, é mais do que duas vezes mais frequente entre negros e mulatos do que entre brancos.

Em suma, a construção do ISS dos sobrenomes nos revelou a existência de um ordenamento socioeconômico de sobrenomes bem definido. No topo da distribuição, estão os sobrenomes de imigrantes recentes, que chegaram ao país durante a segunda metade do século XIX e primeira metade do XX. No estado de São Paulo, ter um sobrenome italiano, sírio-libanes, alemão ou japonês está correlacionado com maior status socioeconômico (na medida que a participação em escolas particulares está correlacionada com maior status).

Os sobrenomes de origem portuguesa, em geral, não apresentam uma padrão definido de status social e racial. Tais sobrenomes estão presentes de forma equânime nas diferentes camadas da população. Contudo, sobrenomes de massa como Silva, Santos e Souza tem baixo status social e são mais negros, assim como os sobrenomes católicos devocionais.

5. Nomes, Sobrenomes e Desempenho Escolar

Nas seções anteriores, estudamos como nomes e sobrenomes estão relacionados ao status social e à raça dos estudantes. Vimos, no caso dos nomes, como pais de diferentes status sociais e níveis educacionais escolhem primeiros nomes diferentes para os filhos. Descobrimos, também, como as medidas de sobrenome contam um pouco sobre a história familiar de longo prazo dos alunos. Seriam nomes e sobrenomes apenas indicadores de diferenças culturais, as quais não afetariam os resultados escolares dos jovens? Ou tais diferenças também afetariam a performance acadêmica? No primeiro cenário, temos que traços culturais revelados pelos nomes apenas indicam condições familiares mas não contribuem para a perpetuação de tais condições. No segundo, tais traços teriam um papel na formação de capital humano das jovens gerações e, portanto, na manutenção do status social familiar. Nosso objetivo nesta seção é avaliar qual dessas duas conjunturas é consistente com os dados.

Com base no exame SARESP 2008, estudamos se nossas medidas sociais e raciais de nomes e sobrenomes conseguem prever proficiência escolar em Português, Matemática e Ciências de alunos de quarta, sexta e oitava séries do ensino fundamental e de terceira série do ensino médio da rede pública estadual. Tal exame é corrigido sem a interferência humana, o que nos garante que nossa medida de performance é imune a um possível viés de correção. Além disso, avaliamos o impacto de nomes e sobrenomes sobre as notas dos boletins escolares de 2008 de Português, Matemática, Ciências, Artes e Educação Física. Para cada medida de performance, rodamos quatro especificações, com ordem crescente

de controles. Os controles são provenientes do questinário socioeconômico do SARESP 2008, o qual inclui indicadores de educação e raça dos pais, renda, raça do aluno e medidas de infraestrutura domiciliar e posse de bens duráveis. Usamos sete variáveis explicativas para caracterizar o nome completo dos alunos, são elas: Índice Social do Nome (ISN), Índice Racial do Nome (IRN), Índice Social do Sobrenome (ISS) herdado do pai e do herdado da mãe, Índice Racial do Sobrenome (IRS) herdado do pai e do herdado da mãe e número de sobrenomes.

Todas as medidas de performance e todos os índices foram normalizados, de forma a facilitar a interpretação dos resultados. Para termos uma idéia das magnitudes envolvidas, o nome Pedro está a quatro desvios-padrão do nome Maikon Douglas no ISN e a três no IRN. O mesmo vale na comparação entre Luiza e Crislane. O nome de família dos Santos, por sua vez, está a três desvios-padrão no ISS e no IRS abaixo de Schmidt, por exemplo.

As tabelas 16, 17 e 18 apresentam os resultados das regressões que tem como variável dependente as notas obtidas no exame SARESP 2008. A primeira coluna de cada tabela tem como controles somente gênero e ano de nascimento do aluno. A idéia dessa especificação é avaliar o poder dos nomes como preditores de performance. Um aumento conjunto de um desvio-padrão em cada um dos índices está associado a elevação de .2825, .2924 e .2868 desvios-padrão em Português, Matemática e Ciências. Considerando que o gap entre meninos e meninas é de cerca de .33 desvios-padrão em Português, temos que nossos índices de nomes e sobrenomes tem um elevado poder preditivo sobre desempenho escolar. Isso evidencia como nomes e sobrenomes captam de maneira substancial características acadêmicas dos alunos. Controlando pelos índices, o número de sobrenomes tem um poder preditivo tímido.

O fato de nomes e sobrenomes serem preditores tão fortes de proficiência escolar é relevante porque nomes completos constam em currículos profissionais. Se uma informação como nome consegue predizer elementos não-observáveis no currículo, há razão para os empregadores usarem nomes e sobrenomes quando das decisões de seleção de currículos para entrevistas de emprego, em um mecanismo de discriminação estatística.

Quando indicadores de raça dos alunos são incluídos como controles, os coeficientes dos índices raciais de nome e sobrenome diminuem sensivelmente. O mesmo não ocorre para os índices sociais, revelando que estes não possuem correlação parcial com a raça reportada. Isso vem a reforçar a análise descritiva dos primeiros nomes, onde mostramos que a relação entre nomes e raça é bem menos pronunciada do que a associação entre nomes e classe social. Por outro lado, dada a alta correlação entre ISS e o IRS é de se esperar que sejam diminutos os coefientes de raça numa regressão do ISS contra IRS e raça.

A terceira especificação inclui o efeito fixo de sala de aula. Se alunos de nomes e sobrenomes tipicamente pobres e negros estão em escolas de pior qualidade, tal controle faz-se necessário. Além disso, se a alocação de alunos entre salas respeitar a ordem alfabética, teremos uma correlação entre nossos índices e a sala de aula. Quando comparamos alunos que estão numa mesma sala de aula, os efeitos de nomes e sobrenomes caem. Se os alunos estudarem em escolas públicas mais próximas de suas casas, e se escolas em vizinhanças menos pobres são de maior qualidade, esse resultado implica que o aluno com nome e/ou sobrenome mais pobre e mais negro tem maiores chances de morar em periferias.

Nossa quarta especificação usa como controles todas as informações do aluno que constam no questionário socioeconômico do SARESP 2008. Entre os controles estão educação e raça dos pais, faixa de renda, infraestrutura domiciliar, posse de bens duráveis e posse de bens culturais (jornais, revistas, internet e dicionário). Nesse cenário, três importantes

padrões se revelam.

O primeiro é a menor importância do IRN perante o ISN. A categoria mais associada a desempenho escolar, portanto, é a de nome de rico ou de pobre e não a de nome de branco ou de negro. Isso é consistente com a noção de que as diferenças culturais responsáveis pelas diferenças nos primeiros nomes são predominantemente determinandas por diferenças sociais e não raciais. A diferença entre Pedro e Jackson ou entre Luisa e Karolaine é da ordem de .09, .078 e .085 desvios-padrão em Português, Matemática e Ciências, respectivamente. Para efeitos de comparação, o gap entre brancos e mulatos está em torno de .04 desvios-padrão; entre brancos e negros, .14.

O segundo deles é o de que a dimensão racial do sobrenome importa mais que a social. O efeito conjunto de uma variação de um desvio-padrão no IRS do pai e da mãe é de mais .02, .029 e .022 desvios-padrão em Português, Matemática e Ciências, respectivamente. Uma variação equivalente no ISS só produziria .0083, .0133 e .0099. O que importa, pois então, não é ter um sobrenome rico, mas ter um sobrenome branco. Não se trata de um quadro onde os sobrenomes importam porque tem a ver com renda e/ou riqueza. Se o padrão racial do sobrenome é que tem relevância, o que importa são traços culturais associados com a etnia e transmitidos entre gerações. Os eventos históricos que geraram o padrão racial de sobrenomes observado atualmente diferenciaram os indivíduos em uma dimensão que até hoje impacta no acúmulo de capital humano destes.

O terceiro padrão é o do sobrenome paterno importar mais que o materno na performance escolar quando medidos na dimensão racial, e o oposto valer quando medidos na dimensão social. Se o IRS capta essencialmente traços culturais familiares, temos evidência de que a cultura trazida pelo pai tem mais efeito sobre performance do que a trazida pela mãe. Por outro lado, o fato do status social do sobrenome das mães ter importância é consis-

tente com um cenário em que investimentos que só podem ser feitos por mães (cuidados na gravidez e amamentação, por exemplo) geram retorno acadêmico futuro.

Nossa última caracterização dos nomes utilizada nas regressões de performance é o número de sobrenomes. Os dados apontam que o efeito de ter um sobrenome a mais está associado a .01 desvios-padrão a mais em Português e Ciências. O efeito é insignificante para Matemática. Testamos ainda se o efeito do ISN é diferente entre Português e Matemática. Se o status social do nome estiver mais associado a habilidades linguísticas dos pais, é de se esperar que o nome seja mais importante em disciplinas intensivas em tais habilidades. Estimamos as equações das duas matérias através de um SUR e realizamos o teste de restrição de parâmetros. Rejeitamos fortemente (p-valor $< .01$) a hipótese de igualdade de parâmetros. Isso indica que alunos com nomes pobres desempenham ainda pior em Língua Portuguesa que em Matemática.

Os resultados acima mostram que efetivamente existe uma relação robusta entre nomes, sobrenomes e desempenho escolar. Como as medidas de performance dos alunos usadas são geradas sem a interferência humana, a hipótese de que tal associação é gerada por viés de correção está descartada.

Nas tabelas 19, 20 e 21, repetimos os exercícios anteriores, mas alterando a medida de proficiência. Em vez de utilizarmos a avaliação do SARESP 2008, nossa variável dependente agora é a nota final do boletim de cada matéria. A nota do boletim se diferencia da do SARESP em basicamente dois aspectos. O primeiro deles é que ela é resultado de avaliações produzidas e corrigidas por professores que conhecem os alunos. O boletim pode refletir, portanto, características dos alunos que não são diretamente medidas em exames padronizados anônimos (comportamento, empenho e habilidades verbais, por exemplo), além de potencialmente captar um efeito de discriminação. O segundo aspecto é

o de que são as notas do boletim as que determinam a progressão do aluno na escola e as que chegam ao conhecimento dos pais.

Os resultados não se alteram substancialmente. Usando o boletim como medida de performance, também encontramos uma relação robusta entre ter nome mais pobre e sobrenome mais negro e pior desempenho. Somente para Ciências observamos uma alteração relevante. Enquanto na medida anônima é o ISN que tem impacto, na nota do boletim, o ISN é insignificante e o IRN é o relevante. Para todas as matérias analisadas, os coeficientes das medidas sociais do sobrenome não são significantes para explicar notas do boletim, reiterando a maior importância relativa da dimensão racial dos nomes de família.

Por último, avaliamos se nomes e sobrenomes também predizem desempenho em matérias mais dependentes de comportamento e participação em classe. As tabelas 22 e 23 apresentam os coeficientes das medidas de primeiro nome e nome de família em regressões que explicam notas em Educação Física e Artes. Na primeira disciplina, só o ISN apresenta efeito estatisticamente significante, enquanto na última, tanto o ISN quanto o IRS do sobrenome herdado do pai têm efeito.

As evidências apresentadas acima são compatíveis com dois cenários distintos. O principal deles é o de seleção em não-observáveis. O segundo é o de que nomes e sobrenomes possuem efeito causal sobre performance escolar. Nossos modelos empíricos são incapazes de separar essas duas hipóteses.

Assim como nomes e sobrenomes estão correlacionados com características observáveis das famílias dos alunos, é provável que também estejam associados com não-observáveis. O pai que escolhe o nome Bernardo, por exemplo, pode ser mais culto que aquele que escolhe Michael Douglas. Famílias com sobrenomes tipicamente brancos podem valorizar

mais a educação que outras (vide exemplo a importância do capital humano em famílias de origem japonesa ou judaica, por exemplo). Nesse mundo, não é o nome em si que causa melhores notas. Nomes e sobrenomes seriam apenas indicadores de universos culturais que, estes sim, causam performance escolar.

Nomes e sobrenomes teriam efeito causal sobre performance se despertassem algum tratamento diferenciado entre professores e alunos. Se professores usam nomes e sobrenomes para formar suas expectativas de proficiência dos alunos, esperando menos de alunos cujos nomes soam mais pobres e mais de alunos com nomes mais ricos, e tais expectativas, positivas ou negativas, se traduzem em diferentes tratamentos, é possível que estudantes desempenhem melhor ou pior simplesmente porque deles se espera isso, num mecanismo de profecias que se auto-realizam. Tal efeito é conhecido na literatura de Psicologia como Pygmalion effect ou Efeito de Rosenthal (ver Rosenthal e Jacobson (1992)).

6. Conclusão

Estudamos nesta dissertação o caráter social e racial de nomes e sobrenomes e sua relação com medidas de acúmulo de capital humano. Os dados indicam que nos anos 90, no Estado de São Paulo, primeiros nomes evoluíram como fortes preditores de status social e educação dos pais. Em especial, nomes pouco comuns e inspirados na língua inglesa estão associados a baixo status. Isso revela como a cultura de massa norte-americana, ao se globalizar, alterou as preferências culturais de brasileiros mais pobres e/ou menos educados. Por outro lado, nomes mais comuns e de origem latina são sinal de maior status. A relação entre primeiros nomes e raça não é tão pervasiva quanto a entre nomes e status social. Quando comparamos jovens de baixo status, não encontramos diferenças relevantes no padrão de nomes entre brancos e afrodescendentes. Isso está de acordo com a noção de que, no Brasil, escolhas culturais são determinadas mais por classe social do que por raça.

Descobrimos ainda que em São Paulo ter um sobrenome não-português está fortemente associado a ter maior status social e ser branco. Nomes de massa, como Silva e Santos, são mais típicos entre pobres e, no caso do último, entre afrodescendentes. Os nomes de família associados ao passado escravocrata (sobrenomes devocionais) apresentam ainda hoje baixos índices sociais. Sobrenomes portugueses, em geral, estão democraticamente distribuídos entre as camadas sociais e raciais da população. Tudo indica que tal cenário foi forjado pelos eventos históricos do final do século XIX (fim da escravidão e imigração em massa) e pela prática, entre portugueses, de repassar seus sobrenomes para seus escravos. Trata-se de uma evidência a favor da hipótese de permanência de status social entre gerações.

Nossas medidas de nomes e sobrenomes apresentam uma relação robusta com desempenho escolar. Alunos com nomes mais pobres (ricos) e sobrenomes mais negros (brancos) apresentam pior performance acadêmica em todas as medidas consideradas. As diferenças de desempenho associadas a nome e sobrenome são altas, especialmente quando comparadas às diferenças raciais.

Nossa primeira conclusão se refere a capacidade de predição que os primeiros nomes tem. Podemos afirmar que a relação entre nomes e status socioeconômico é um característica marcante da atual geração de jovens nascidos entre o final dos anos 80 e começo dos 2000. Isso reflete um cenário em que a desigualdade socioeconômica está associada a outra desigualdade, a de capital cultural familiar. Os pais que escolhem o nome Jackson para o filho (nome fictício) são, em média, mais pobres e menos educados do que os pais de um Heitor, por exemplo. Provavelmente o nome Jackson foi gerado por um padrão cultural muito influenciado pelas linguagens de massa (televisão e cinema), enquanto Heitor por um padrão mais influenciado por linguagem escrita. Tais diferenciais culturais familiares

se refletem nos primeiros nomes e, mais importante, são transmitidos para os filhos. Como resultado, Heitor tem, em média, notas melhores que as de Jackson.

Nossa segunda conclusão diz respeito aos sobrenomes. Se há 100 anos a família Santos era, em média, mais pobre e mais negra que a família Rossi (sobrenomes fictícios), esse quadro ainda se mantém nos dias de hoje. Mais importante que isso, o filho da família Santos tem, em média, menores notas do que o filho da família Rossi, apontando para continuidade da desigualdade no futuro. Ao contrário do nome, que reflete a cultura da família no curto prazo, temos aqui um relação entre um traço cultural que acompanha o sobrenome entre as gerações e o desempenho acadêmico das atuais gerações. Certamente, o estudo mais profundo de tal relação, e a identificação precisa desses elementos culturais, são caminhos abertos por nossos resultados.

7. Apêndice

Tabela 21: Proficiência em Português no exame SARESP 2008

ISN	.0679	.0683	.0457	.0225
	(.0010)	(.0010)	(.0010)	(.0010)
IRN	.0434	.0194	.0119	.0037
	(.0010)	(.0010)	(.0010)	(.0010)
$ISS_{mãe}$	.0266	.0264	.0140	.0046
	(.0012)	(.0012)	(.0011)	(.0011)
ISS_{pai}	.0266	.0258	.0120	.0037
	(.0012)	(.0012)	(.0012)	(.0012)
$IRS_{mãe}$	.0523	.0402	.0206	.0085
	(.0012)	(.0012)	(.0012)	(.0011)
IRS_{pai}	.0657	.0505	.0251	.0119
	(.0013)	(.0013)	(.0012)	(.0012)
N. de Sobrenomes	.0072	.0102	.0226	.0102
	(.0019)	(.0019)	(.0018)	(.0017)
N. de Observações	980.917	980.917	980.917	980.917
Raça	Não	Sim	Sim	Sim
Sala de Aula	Não	Não	Sim	Sim
Socioeconômico	Não	Não	Não	Sim

Erros-padrão corrigidos por cluster de sala.

Todos os coeficientes são significantes a 1%.

Tabela 22: Proficiência em Matemática no exame SARESP 2008

ISN	.0606	.0609	.0414	.0197
	(.0011)	(.0011)	(.0010)	(.0010)
IRN	.0410	.01762	.0091	.0012*
	(.0010)	(.0010)	(.0010)	(.0010)
$ISS_{mãe}$	.0288	.0284	.0160	.0070
	(.0012)	(.0012)	(.0012)	(.0011)
ISS_{pai}	.0297	.0288	.014	.0063
	(.0013)	(.0013)	(.0012)	(.0012)
$IRS_{mãe}$	.0580	.0457	.0244	.0120
	(.0012)	(.0012)	(.0012)	(.0012)
IRS_{pai}	.0743	.058	.0311	.0173
	(.0013)	(.0013)	(.0013)	(.0013)
N. de Sobrenomes	-.0089	-.0063	.0141	.0030**
	(.0019)	(.0019)	(.0018)	(.0018)
N. de Observações	980.917	980.917	980.917	980.917
Raça	Não	Sim	Sim	Sim
Sala de Aula	Não	Não	Sim	Sim
Socioeconômico	Não	Não	Não	Sim

Erros-padrão corrigidos por cluster de sala.

Todos os coeficientes são significantes a 1%, salvo:** sig. a 10%; * não sig.

Tabela 23: Proficiência em Ciências no exame SARESP 2008

ISN	.0635	.0636	.0435	.0213
	(.0011)	(.0011)	(.0011)	(.0011)
IRN	.0438	.0189	.0106	.0023*
	(.0011)	(.0011)	(.0011)	(.0011)
$ISS_{mãe}$	.0264	.0260	.0141	.0051
	(.0013)	(.0013)	(.0013)	(.0012)
ISS_{pai}	.0277	.0267	.0129	.0048
	(.0013)	(.0013)	(.0013)	(.0013)
$IRS_{mãe}$	.0544	.0416	.0218	.0097
	(.0013)	(.0013)	(.0013)	(.0013)
IRS_{pai}	.0710	.0548	.0287	.0126
	(.0014)	(.0014)	(.0013)	(.0013)
N. de Sobrenomes	.0041*	.0074	.0241	.012
	(.0020)	(.0020)	(.0020)	(.0019)
N. de Observações	850,865	850,865	850,865	850,865
Raça	Não	Sim	Sim	Sim
Sala de Aula	Não	Não	Sim	Sim
Socioeconômico	Não	Não	Não	Sim

Erros-padrão corrigidos por cluster de sala

Todos os coeficientes são significantes a 1%, salvo: * sig. a 5%

Tabela 24: Nota média no boletim de Português de 2008

ISN	.0404	.0408	.0187
	(.0010)	(.0010)	(.0010)
IRN	.0240	.0071	-.0007*
	(.0010)	(.0010)	(.0010)
$ISS_{mãe}$	.0088	.0090	.0009*
	(.0012)	(.0012)	(.0011)
ISS_{pai}	.0074	.0073	.0003*
	(.0012)	(.0012)	(.0012)
$IRS_{mãe}$	.0290	.0212	.0102
	(.0012)	(.0012)	(.0012)
IRS_{pai}	.0359	.0261	.0137
	(.0013)	(.0013)	(.0012)
N. de Sobrenomes	.0096	.0100	.0010*
	(.0018)	(.0018)	(.0018)
N. de Observações	980.718	980.718	980.718
Raça	Não	Sim	Sim
Sala de Aula	Sim	Sim	Sim
Socioeconômico	Não	Não	Sim

Erros-padrão corrigidos por cluster de sala.

Todos os coeficientes são significantes a 1%, salvo: * não sig.

Tabela 25: Nota média no boletim de Matemática de 2008

ISN	.0380	.0384	.0163
	(.0011)	(.0011)	(.0010)
IRN	.0224	.0053	-.0026**
	(.0010)	(.0011)	(.0010)
$ISS_{mãe}$	.0112	.0112	.0029**
	(.0012)	(.0012)	(.0012)
ISS_{pai}	.0088	.0086	.0012*
	(.0013)	(.0013)	(.0013)
$IRS_{mãe}$	.0324	.0243	.0125
	(.0013)	(.0013)	(.0013)
IRS_{pai}	.0411	.0308	.0174
	(.0013)	(.0013)	(.0013)
N. de Sobrenomes	.0092	.0095	.0005*
	(.0019)	(.0019)	(.0018)
N. de Observações	980.721	980.721	980.721
Raça	Não	Sim	Sim
Sala de Aula	Sim	Sim	Sim
Socioeconômico	Não	Não	Sim

Erros-padrão corrigidos por cluster de sala.

Todos os coeficientes são significantes a 1%, salvo: * não sig.; ** sig. a 5%.

Tabela 26: Nota média no boletim de Ciências de 2008

ISN	.0428	.0434	.0003*
	(.0012)	(.0012)	(.0012)
IRN	.0227	.0042	.0207
	(.0012)	(.0012)	(.0012)
$ISS_{mãe}$	.0112	.0113	.0021*
	(.0014)	(.0014)	(.0014)
ISS_{pai}	.0092	.0091	.0009*
	(.0015)	(.0015)	(.0015)
$IRS_{mãe}$	.0299	.0207	.0078
	(.0015)	(.0015)	(.0015)
IRS_{pai}	.0391	.0276	.0124
	(.0016)	(.0016)	(.0016)
N. de Sobrenomes	.0142	.0145	.0103
	(.0022)	(.0022)	(.0022)
N. de Observações	739.666	739.666	739.666
Raça	Não	Sim	Sim
Sala de Aula	Sim	Sim	Sim
Socioeconômico	Não	Não	Sim

Erros-padrão corrigidos por cluster de sala.

Todos os coeficientes são significantes a 1%, salvo: * não sig;** sig a 5%.

Tabela 27: Nota média no boletim de Educação Física de 2008

ISN	.0288	.0289	.0132
	(.0012)	(.0012)	(.0011)
IRN	.0093	.0026*	-.0025**
	(.0011)	(.0012)	(.0011)
$ISS_{mãe}$	.0064	.0064	.0012*
	(.0013)	(.0013)	(.0013)
ISS_{pai}	.0056	.0055	.0009*
	(.0014)	(.0014)	(.0014)
$IRS_{mãe}$	.0134	.0104	.0036
	(.0013)	(.0013)	(.0013)
IRS_{pai}	.0175	.0138	.0056
	(.0014)	(.0014)	(.0014)
N. de Sobrenomes	.0061*	.0064	.0001*
	(.0021)	(.0021)	(.0021)
N. de Observações	867.426	867.426	867.426
Raça	Não	Sim	Sim
Sala de Aula	Sim	Sim	Sim
Socioeconômico	Não	Não	Sim

Erros-padrão corrigidos por cluster de sala

Todos os coeficientes são significantes a 1%, salvo: * não sig; ** sig a 5%.

Tabela 28: Nota média no boletim de Artes de 2008

ISN	.0296	.0300	.0121
	(.0012)	(.0012)	(.0012)
IRN	.0167	.0018*	-.0035**
	(.0012)	(.0012)	(.0012)
$ISS_{mãe}$	.0070	.0070	.0014*
	(.0014)	(.0014)	(.0014)
ISS_{pai}	.0030**	.0029**	-.002*
	(.0015)	(.0015)	(.0015)
$IRS_{mãe}$	.0214	.0139	.0054
	(.0014)	(.0014)	(.0014)
IRS_{pai}	.0298	.0204	.0104
	(.0016)	(.0016)	(.0015)
N. de Sobrenomes	.0059	.0061	-.0010*
	(.0022)	(.0022)	(.0022)
N. de Observações	735.200	735.200	735.200
Raça	Não	Sim	Sim
Sala de Aula	Sim	Sim	Sim
Socioeconômico	Não	Não	Sim

Erros-padrão corrigidos por cluster de sala.

Todos os coeficientes são significantes a 1%, salvo: * não sig.; ** sig. a 5%.

Tabela 29: Estatísticas Descritivas

Variável	Média	Desvio-Padrão
ISN	39.59	20.93
IRN	50.61	17.28
ISS	46.52	13.33
IRS	50.35	13.74
Feminino	.523	.499
Branco	.492	.499
Negro	.084	.277
Mulato	.372	.483
Indígena	.0187	.135
Asiático	.020	.142
Não declarou Raça	.01	.10
Ano de Nascimento	1993	2.5
Pessoas no Domicílio	4.45	1.37
Uma TV	.415	.492
Duas TVs	.526	.499
Um DVD	.685	.464
Dois DVDs	.161	.3681
Um PC	.46	.498
Dois PCs	.042	.2
Uma Máquina de Lavar	.702	.456
Duas Máquinas de Lavar	.059	.236
Um carro	.41	.491
Dois carros	.076	.265
Um banheiro	.643	.478
Dois banheiros	.263	.44
Gás encanado	.205	.404
Água Encanada	.959	.198
Luz Elétrica	.957	.202
Jornais e/ou Revistas	.794	.403
Dicionário	.89	.311
Internet	.37	.483
Tempo translado casa/escola	33	11.45
Casa própria paga	.546	.497
Casa própria financiada	.106	.307
Casa alugada	.188	.39
Renda 1: Até 850	.283	.45
Renda 2: 851 a 1275	.246	.43
Renda 3: 1276 a 2125	.173	.378
Renda 4: 2126 a 4250	.067	.25
Renda 5: acima de 4250	.02	.14
Não declarou renda	.209	.406
Mãe - Sem escola	.072	.259
Mãe - Primeiro Grau	.421	.493
Mãe - Segundo Grau	.158	.365
Mãe - Ensino Superior	.061	.24
Mãe - Não reportou educação	.285	.451
Pai - Sem escola	.036	.187
Pai - Primeiro Grau	.446	.497
Pai - Segundo Grau	.173	.378
Pai - Ensino Superior	.065	.247
Pai - Não reportou educação	.278	.448

8. Referências Bibliográficas

AIGNER, Dennis J; CAIN, Glen G. *Statistical theories of discrimination in labor markets.* **Industrial and Labor Relations Review**. Vol. 30, n.2, p. 175-187, 1977.

AKERLOF, George A; KRANTON, Rachel E. *Economics and Identity.* **The Quartely Journal of Economics**. Vol. 115, n.3, p. 715-753, 2000.

AKERLOF, George A; KRANTON, Rachel E. *Identity and Schooling: Some lessons for the Economics of Education.* **Journal of Economic Literature**. Vol. 40, n.4, p. 1167-1201, 2002.

ARAI, Mahmood; Thoursie, Peter S. *Renouncing Personal Names: An Empirical Examination of Surname Change and Earnings.* **Journal of Labor Economics**. Vol. 27, n.1, p. 127-147, 2009.

ARAI, Mahmood et al. *Childrens first names and immigration background in France.* **SULCIS Working Papers, Stockholm University Linnaes Center for Integration Studies, 2009**. Maio de 2009.

AZEVEDO, Eliane S. *The anthropological and cultural meaning of family names in Bahia.* **Current Anthropology**. Vol. 21, n.1, p. 360-363, 1980.

BAENINGER, Rosana.**Atlas da Imigração Internacional em São Paulo: 1850-1950**. São Paulo, SP: Editora da Universidade de São Paulo, 2009.

BARBOSA, Ana et al. *Family names and the length of the Y chromosome in Brazilian blacks.* **Brazilian Journal of Genetics**. Vol. 20, n.1, p. 93-96, 1997.

BERTRAND, Marianne; MULLAINATHAN, Sendhil. *Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination.* **American Economic Review**. Vol. 94, n.4, p. 991-1013, 2004.

COLLADO, Dolores; ORTÍN, Ignacio O.;ROMEU, Andrés. *Surnames and Social Status in Spain.* **Investigaciones Economicas**. Vol. 32, n.3, p. 83-116, 2008.

FIGLIO, David N. *Names, Expectations and Black-White Test Score Gap.* **NBER Working Paper Series**. N. 11195, Março de 2005.

FIGLIO, David N. *Boys Named Sue: Disruptive Children and their Peers.* **NBER Working Paper Series**. N. 11277, Abril de 2005.

FRYER, Roland; LEVITT, Steven. *The causes and consequences of distinctively Black Names.* **The Quartely Journal Of Economics**. Vol. 119, n.3, p. 767-805, 2004.

GREENE, William. **Econometric Analysis**. 6th ed. Upper Saddle River, NJ: Prentice Hall, 2007.

JOBLING, Mark A. *In the name of the father: surnames and genetics.* **Trends in Genetics**. Vol. 17, n. 6, p. 353-357, 2001.

LA FERRARA, Eliana; Chong, Alberto; Duryea, Suzanne. *Soap Operas and Fertility: Evidence from Brazil.* **BREAD Working Paper**. N. 172, Março de 2008.

LAVY, Victor. *Do Gender stereotypes reduce girls or boys human capital outcomes? Evidence from a natural experiment.* **Journal of Public Economics**. Vol. 92, p. 2083-2105, 2008.

MADEIRA, Ricardo A.; Vicentini, Gustavo. *The Median Ideal Policy in Mixed Markets for Education: Evidence from Brazil.* **The Northeast Universities Development Consortium Conference - NEUDC 2006, 2006, Ithaca.**

MOURA, Clóvis. **Dicionário da Escravidão Negra no Brasil**. São Paulo, SP: Editora da Universidade de São Paulo, 2004.

PHELPS, Edmund S. *The statistical theory of racism and sexism.* **American Economic Review**. Vol. 62, n.4, p. 659-661, 1972.

ROSENTHAL, Robert; JACOBSON, Lenore. **Pygmalion in the classroom**. Expanded edition. New York: Irvington, 1992.