

Victor Haselmann Arakawa

**Concluintes do Ensino Médio: a
Qualidade da Educação e a Situação
Ocupacional dos Jovens no Brasil
Metropolitano**

Belo Horizonte, MG
UFMG/Cedeplar
2012

Victor Haselmann Arakawa

Concluintes do Ensino Médio: a Qualidade da Educação e a Situação Ocupacional dos Jovens no Brasil Metropolitano

Dissertação apresentada ao curso de mestrado em economia do Centro de Desenvolvimento e Planejamento Regional da Faculdade de Ciências Econômicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do Título de Mestre em Economia.

Orientador: Prof. Dr. André Braz Golgher

Co-orientadora: Prof^a. Dr^a Ana Flávia Machado

Belo Horizonte, MG
Centro de Desenvolvimento e Planejamento Regional
Faculdade de Ciências Econômicas - UFMG
2012

Folha de Aprovação

*Ao meu avô Mansueto Paulo e minha avó
Judith Haselmann Paulo*

AGRADECIMENTOS

Agradeço em primeiro lugar ao Cedeplar, pela oportunidade de participar de um ambiente de conhecimento amplamente reconhecido pela sua qualidade e comprometimento com o avanço teórico e aplicado da ciência econômica.

Agradeço ao colegiado do programa de pós-graduação, por prezar pela solidez do curso e cuidado com todos, incluindo alunos, professores e pesquisadores. A CAPES também merece reconhecimento pela bolsa concedida, que possibilitou minha mudança de São Paulo para Belo Horizonte, com dedicação exclusiva ao curso.

Os professores ligados à coordenação do curso foram atenciosos durante todo o meu contato com a parte institucional e, por isso, gostaria de agradecer ao apoio de Frederico Gonzaga Jayme Junior e Ana Maria Hermeto Camilo de Oliveira. Agradeço a Rodrigo Ferreira Simões pela atenção institucional durante a época em que fui monitor do curso de relações econômicas internacionais e professor substituto de microeconomia. A oportunidade de poder lecionar em uma instituição como a UFMG é algo único e que certamente me recordarei positivamente.

Sou muito agradecido a todos os professores que se dedicaram para transmitir o seu conhecimento aos alunos. Gostaria de lembrar dos professores Cássio Maldonado Turra e Simone Wajnman, do programa de pós-graduação em Demografia, pelas aulas de demografia econômica que foram fundamentais para a discussão deste trabalho.

Agradeço a professora Ana Maria Hermeto Camilo de Oliveira pela atenção fora da sala de aula, para orientação sobre o rumo acadêmico dos alunos, e por aceitar o convite para fazer parte da banca de defesa desta dissertação. Da mesma maneira, agradeço Sergei Soares por também aceitar o convite para participar da banca examinadora.

Agradeço enormemente à minha coorientadora Ana Flávia Machado, pelo interesse e comentários sobre a ideia principal sobre educação, tão ampla e

interessante, bem como a orientação sobre a manipulação da base de dados da PME, que pode ser um trabalho homérico.

Ao meu orientador, André Braz Golgher, agradeço pela enorme disponibilidade em discutir todos os pontos do trabalho, sugerindo alternativas criativas para problemas nem um pouco triviais. A oportunidade de trabalhar junto, também como seu monitor, foi uma boa experiência que certamente contribuiu para a minha formação.

A todos aqueles que compartilharam os desafios do nivelamento e do primeiro semestre, gostaria de agradecer pela amizade que certamente levarei de Minas Gerais. Obrigado aos colegas de mestrado e de doutorado.

Agradeço à minha família, pela educação que me deu e pelo apoio na mudança de São Paulo. Por fim, agradeço à minha esposa Luzia Maria Cavalcante de Melo, que amo e que conheci no CEDEPLAR e que foi compreensiva com a dedicação de grande parte do nosso tempo à conclusão deste trabalho.

SUMÁRIO

1 INTRODUÇÃO	14
2 REVISÃO BIBLIOGRÁFICA.....	17
2.1 Sobre a Teoria do Capital Humano e suas Críticas	17
2.2 Fatores Associados ao Desempenho Escolar.....	20
2.2.1 Características Pessoais	23
2.2.2 Nível Socioeconômico	26
2.2.3 Estrutura e Arranjo Familiar.....	27
2.3 Determinantes Escolares da Proficiência Cognitiva	29
2.4 Determinantes Ocupacionais pós-Ensino Médio.....	31
3 METODOLOGIA.....	35
3.1 Estratégia Empírica	37
3.1.1 Estimação da Função de Produção Educacional	37
3.1.2 A Correção do Viés de Auto-Seleção.....	38
3.1.2.1 Metodologia da Correção do Viés de Auto-Seleção.....	38
3.1.3 A Violação da Hipótese de Amostras Independentes e Identicamente Distribuídas	42
3.1.3.1 Metodologia dos Modelos Hierárquicos Lineares.....	43
3.1.4 Imputação da Nota Estimada do ENEM na PME	45
3.1.5 Estimação da Situação Ocupacional do Indivíduo em Função da Proficiência	48
3.2 Apresentação das Bases de Dados e Variáveis Utilizadas na Estratégia Empírica	51
3.2.1 Descrição da Base de Dados do ENEM e das Variáveis da Função de Produção Educacional	51
3.2.2 Descrição da Base de Dados do Censo Escolar e das Variáveis Utilizadas na Correção de Heckman	56

3.2.3 Descrição da Base de Dados da PME Utilizada para a Estimação da Situação Ocupacional dos Indivíduos	57
4 ANÁLISE DESCRITIVA DAS VARIÁVEIS UTILIZADAS NA ESTRATÉGIA EMPÍRICA.....	61
4.1 Análise Descritiva das Variáveis Utilizadas na Estimação da Função de Produção Educacional	61
4.2 Análise Descritiva das Variáveis Escolares utilizadas na Correção do Viés de Auto-Seleção.....	68
4.3 Análise Descritiva das Variáveis Utilizadas na Imputação da Proficiência Cognitiva do ENEM para a PME	71
4.4 Análise Descritiva da Variável de Situação Ocupacional na Base de Dados da PME	78
5. FUNÇÃO DE PRODUÇÃO EDUCACIONAL.....	80
5.1 Estimação da Função de Produção Educacional por MQO	80
5.2 Função de Produção Educacional com Correção para o Viés de Auto-Seleção do ENEM.....	85
5.3 Estimação da Função de Produção Educacional por meio de Modelos Hierárquicos Lineares	88
6 A RELAÇÃO ENTRE PROFICIÊNCIA COGNITIVA E SITUAÇÃO OCUPACIONAL DOS JOVENS.....	96
6.1 Imputação da Proficiência Cognitiva do ENEM para a PME	96
6.2 Análise dos Determinantes da Mobilidade Ocupacional dos Jovens	99
7 CONCLUSÃO.....	106
REFERÊNCIAS BIBLIOGRÁFICAS	109
ANEXOS	116

LISTA DE ILUSTRAÇÕES

FIGURA 1 - MODELO CONCEITUAL EXPLICATIVO DA PROFICIÊNCIA COGNITIVA	22
QUADRO 1 – ESTUDOS DE PERSISTÊNCIA INTERGERACIONAL DE EDUCAÇÃO.....	25
QUADRO 2 - NÚMERO DE INDIVÍDUOS POR MÊS E ANO DE ENTRADA NA PME	60
TABELA 1 – RESUMO DAS ESTATÍSTICAS DESCRITIVAS DE TODAS AS VARIÁVEIS UTILIZADAS NA ESTIMAÇÃO DA FUNÇÃO DE PRODUÇÃO EDUCACIONAL	62
GRÁFICO 1 - DISTRIBUIÇÃO ACUMULADA DA IDADE DOS CANDIDATOS POR REDE DE ENSINO.....	63
GRÁFICO 2 - DISTRIBUIÇÃO ACUMULADA DA IDADE DOS CANDIDATOS QUE ESTÃO NO 3º ANO DO ENSINO MÉDIO SEGUNDO A ESCOLARIDADE DA MÃE.	64
TABELA 2 – DISTRIBUIÇÃO TOTAL E POR REGIÃO METROPOLITANA DOS CANDIDATOS SEGUNDO A RENDA FAMILIAR	65
GRÁFICO 3 - PROPORÇÃO DE CANDIDATOS AO ENEM SEGUNDO GÊNERO E FAIXA DE RENDA FAMILIAR.....	66
TABELA 3 – DISTRIBUIÇÃO DOS CANDIDATOS POR RM SEGUNDO ESCOLARIDADE DA MÃE	67
QUADRO 3 - ESTATÍSTICAS DESCRITIVAS SOBRE GÊNERO E COR/RAÇA POR REGIÃO METROPOLITANA.....	67
TABELA 4 - ESTATÍSTICAS DESCRITIVAS DA IDADE DOS DOCENTES POR REGIÃO METROPOLITANA.....	69
TABELA 5 - ESTATÍSTICAS DESCRITIVAS DA PROPORÇÃO DE DOCENTES COM PÓS-GRADUAÇÃO NAS ESCOLAS POR REGIÃO METROPOLITANA	70

TABELA 6 - PROPORÇÃO TOTAL DAS ESCOLAS POR TIPO DE INFRAESTRUTURA	71
GRÁFICO 4 - DISTRIBUIÇÃO DOS INDIVÍDUOS POR IDADE NA BASE DE DADOS DO ENEM E PME	72
GRÁFICO 5 - DISTRIBUIÇÃO DOS INDIVÍDUOS SEGUNDO ESCOLARIDADE DA MÃE NA BASE DE DADOS DO ENEM E PME	73
GRÁFICO 6 - DISTRIBUIÇÃO PERCENTUAL DOS INDIVÍDUOS SEGUNDO FAIXA DE RENDA FAMILIAR NA BASE DE DADOS DA PME E ENEM	75
GRÁFICO 7 - DISTRIBUIÇÃO PERCENTUAL DOS INDIVÍDUOS SEGUNDO A QUANTIDADE DE MORADORES NO DOMICÍLIO NA BASE DE DADOS DA PME E ENEM	76
GRÁFICO 8 - DISTRIBUIÇÃO PERCENTUAL DOS INDIVÍDUOS SEGUNDO A REGIÃO METROPOLITANA EM QUE MORA	77
QUADRO 4: MATRIZ DE TRANSIÇÃO DA SITUAÇÃO OCUPACIONAL EM DOZE MESES	79
TABELA 7 - RESULTADO DOS MODELOS ECONÔMICOS ESTIMADOS PARA FUNÇÃO DE PRODUÇÃO EDUCACIONAL DA NOTA DA PROVA OBJETIVA NO ENEM 2008.....	94
TABELA 8 - RESULTADO DOS MODELOS ECONÔMICOS ESTIMADOS PARA FUNÇÃO DE PRODUÇÃO EDUCACIONAL DA NOTA DA PROVA DE REDAÇÃO DO ENEM 2008	95
TABELA 9 - RESULTADO DA ESTIMAÇÃO DO MODELO (3.17) PARA A IMPUTAÇÃO DA NOTA DO ENEM PARA OS INDIVÍDUOS DA PME POR IDADE	97
TABELA 10 – ESTATÍSTICAS DESCRITIVAS DA NOTA DO ENEM E DA PROFICIÊNCIA IMPUTADA NA BASES DE DADOS DA PME	98
GRÁFICO 9 - DISTRIBUIÇÃO DA NOTA DO ENEM 2008 E DA NOTA IMPUTADA NA PME 2008-2009.....	98

TABELA 11 - RAZÃO DE RISCO RELATIVO PARA AS SITUAÇÕES OCUPACIONAIS ESTIMADAS PELO MODELO LOGIT MULTINOMIAL PARA A SITUAÇÃO INICIAL "TRABALHA E ESTUDA"	102
TABELA 12 - RAZÃO DE RISCO RELATIVO PARA AS SITUAÇÕES OCUPACIONAIS ESTIMADAS PELO MODELO LOGIT MULTINOMIAL PARA A SITUAÇÃO INICIAL "NÃO TRABALHA E ESTUDA"	103
QUADRO A.1: VARIÁVEIS UTILIZADAS NO PROCESSO DE IMPUTAÇÃO DA NOTA DO ENEM NA PME.....	116
GRÁFICO A.1 - DISTRIBUIÇÃO DO LOG DA RENDA PER CAPITA NA BASE DE DADOS DA PNAD E PME.....	117
GRÁFICO A.2 - DISTRIBUIÇÃO PERCENTUAL DOS INDIVÍDUOS SEGUNDO A REGIÃO METROPOLITANA EM QUE RESIDE, PELA BASE DE DADOS DO ENEM E DA PME COM EXPANSÃO DA AMOSTRA	118
TABELA A.1 - PRIMEIRO ESTÁGIO DE HECKMAN - MODELO PROBIT PARA A PROBABILIDADE DO INDIVÍDUO COMPARECER À PROVA OBJETIVA DO ENEM.....	119
TABELA A.2 - PRIMEIRO ESTÁGIO DE HECKMAN - MODELO PROBIT PARA A PROBABILIDADE DO INDIVÍDUO COMPARECER À PROVA DE REDAÇÃO DO ENEM.....	120
TABELA A.3 - RESULTADO DO MODELO (3.19) QUE ESTIMA O ERRO DO MODELO DE IMPUTAÇÃO DA NOTA DO ENEM NA BASE DA PME POR IDADE	121

RESUMO

Esta dissertação investiga os determinantes da proficiência cognitiva dos estudantes do último ano do ensino médio, nas regiões metropolitanas de Belo Horizonte, São Paulo, Rio de Janeiro, Salvador, Recife e Porto Alegre. Além disso, investiga a importância da proficiência como determinante da situação ocupacional dos jovens após a conclusão da educação básica.

O método primário utilizado para estimar a função de produção educacional, com base no ENEM 2008 e no Censo Escolar 2008, foi o de mínimos quadrados ordinários. A presença do viés de auto-seleção da amostra e do viés de dependência das amostras foram testadas e corrigidas, respectivamente, pelo método em dois estágios de Heckman (1979) e pelo método dos modelos hierárquicos lineares.

A imputação da nota do ENEM 2008 na PME 2008-2009 foi baseada na metodologia de criação de mapas de pobreza. Após a imputação da variável de proficiência, utilizou-se o modelo probabilístico multinomial *logit* para estimar a probabilidade de transição entre situações ocupacionais, tendo a proficiência cognitiva como determinante, utilizando a PME de 2008 a 2010.

Como resultado dos principais determinantes da nota do ENEM em 2008, a renda familiar e o nível educacional da mãe do estudante foram os mais importantes. Além disso, negros/pardos possuem notas menores que brancos/amarelos, as mulheres possuem desempenho inferior que os homens na prova objetiva, mas obtêm melhores resultados na redação. Considerando as diferenças regionais, a região metropolitana de Belo Horizonte possui uma nota média superior que todas as demais.

O resultado do modelo probabilístico para a situação ocupacional dos jovens mostrou que não existe muita diferença entre sexos, mas a proficiência cognitiva aumenta a probabilidade do indivíduo continuar estudando.

Palavras-chave: proficiência cognitiva, imputação de dados, situação ocupacional.

ABSTRACT

This dissertation investigates the determinants of cognitive proficiency of students in their final year of high school in the metropolitan area of Belo Horizonte, São Paulo, Rio de Janeiro, Salvador, Recife and Porto Alegre. Furthermore, it investigates the importance of cognitive proficiency as a determinant of occupational status of young people after completion of basic education.

The primary method used to estimate the educational production function, based on the ENEM and School Census databases of 2008, was the ordinary least squares. The presence of self-selection bias in the sample and the bias of non i.i.d. samples were fixed with, respectively, the method of two stage of Heckman (1979) and the method of hierarchical linear models.

The imputation of ENEM scores in PME was based on the methodology of poverty maps creation. After the imputation of the cognitive proficiency variable, we used the multinomial logit model to estimate the probability of transition between occupational status, with cognitive proficiency as a determinant, using the PME from 2008 to 2010.

As a result of the main determinants of the ENEM score in 2008, we can point family income and educational level of student's mother as the most important. In addition, black students have lower grades than white or asian students, women have lower performance than men in the objective test, but perform better in writing. Considering regional differences, the metropolitan area of Belo Horizonte has an average score higher than all others.

The result of the probabilistic model for the occupation of young people showed that there is not much difference between sexes, but the cognitive proficiency increases the probability of an individual to continue studying in higher education.

Keywords: cognitive proficiency, data mashup, occupational status.

1 INTRODUÇÃO

A influência da educação dos indivíduos sobre as suas oportunidades no mercado de trabalho e sobre como o mercado de trabalho valoriza a educação faz parte de uma ampla discussão teórica e empírica no ramo da economia e outras ciências sociais, sendo esta, a principal motivação para este trabalho.

A maioria dos indivíduos reserva boa parte da juventude dedicada ao sistema educacional formal. Tal sistema deve proporcionar aos indivíduos o desenvolvimento das suas faculdades cognitivas, utilizadas no processamento de informações, relacionadas ao raciocínio lógico e, paralelamente, instigar o desenvolvimento de habilidades não-cognitivas, ligadas principalmente à sociabilidade e sensibilidade humana. Assim, a qualidade da formação educacional pode influenciar a inserção no mercado de trabalho, uma vez que este valoriza tais habilidades pessoais.

Alguns economistas (MINCER, 1958; BECKER, 1962; HANUSCHEK, 2009; HALEY, 1973), por meio da Teoria do Capital Humano, afirmam que a escolaridade afeta a renda dos indivíduos, mas não deixam claro se isso decorre da habilidade cognitiva adquirida na escola, apesar de ser este o argumento subjacente na teoria. Além disso, talvez, o simples fato das pessoas freqüentarem a escola seja tão importante quanto o seu conteúdo propriamente dito (BOWLES et. al, 2001; HECKMAN & RUBINSTEIN, 2001). O ambiente escolar pode proporcionar ao estudante não apenas a capacitação intelectual, mas também desenvolvimento de uma rede de amizades e habilidades sociais, fruto da interação interpessoal.

Nesse contexto, o foco deste trabalho concentra-se no estudo da relação entre as habilidades cognitivas de um indivíduo e a sua posição no mercado de trabalho após a conclusão do ensino médio, medidas, respectivamente, por meio de sua proficiência no Exame Nacional do Ensino Médio (ENEM) de 2008 e a situação ocupacional (somente estuda, somente trabalha, estuda e trabalha ou nenhuma

das opções) inicial e final, na periodicidade de 12 meses, utilizando a Pesquisa Mensal de Emprego (PME) de 2008 a 2010.

Nosso objetivo principal é estudar a relação entre a posição dos indivíduos no mercado de trabalho, doze meses após a conclusão do ensino médio, e a sua proficiência cognitiva. Para isso, imputamos a nota estimada do ENEM para os indivíduos na base de dados da PME de 2008 e 2009, segundo características comuns a ambas as bases.

Para se alcançar este objetivo principal, precisamos, em primeiro lugar, levantar os principais trabalhos relacionados ao tema e, a partir dessa discussão, estimar e analisar a função de produção educacional, utilizando a nota do ENEM 2008 e os seus determinantes (características pessoais, *background* familiar e escolar), que faz parte do nosso objetivo secundário.

As hipóteses principais são de que a renda familiar, a educação dos pais e a frequência de uma escola de melhor qualidade e a não competição entre o tempo de estudos com o trabalho auxiliam na proficiência dos estudantes. Além disso, espera-se que a maior proficiência cognitiva dos recém egressos do ensino médio aumente a probabilidade de que eles busquem dar continuidade aos estudos no ensino superior, pela diminuição do custo de aprendizado.

Nesse caso, a compatibilização da base do ENEM com a PME é fundamental para alcançar os objetivos da dissertação, uma vez que cada uma delas possui uma parte essencial para a investigação empírica da influência da proficiência escolar nas decisões de alocação do tempo dos jovens.

A compatibilização que se busca neste trabalho não é trivial, uma vez que não observamos os mesmos indivíduos nas duas amostras e, também, porque podem existir vieses, que são discutidos nos capítulos subsequentes. Os trabalhos pioneiros (MINCER, 1958; BECKER, 1962) estimaram a relação causal entre escolaridade, medida em anos de estudo, e renda do indivíduo, mas atualmente, a disponibilidade de dados de exames de proficiência abriu novas portas para estudar as habilidades cognitivas, a renda e a posição dos indivíduos no mercado de trabalho.

Esta dissertação está dividida da seguinte forma: no segundo capítulo, é feita uma revisão bibliográfica sobre trabalhos considerados relevantes para os objetivos da análise, tanto em termos teóricos quanto aplicados, com a finalidade de situar nossos objetivos e contribuições. No terceiro capítulo, descrevemos a metodologia, detalhando as bases de dados e as variáveis a serem utilizadas em cada etapa da dissertação. No quarto capítulo, apresentamos a análise descritiva das variáveis que foram utilizadas na estratégia empírica.

No quinto capítulo, estimou-se a função de produção educacional utilizando o método de mínimos quadrados ordinários (MQO), a correção de Heckman para o problema de viés de auto-seleção e os modelos hierárquicos para o problema de independência das amostras. Os determinantes da proficiência cognitiva dos jovens que estão no último ano do ensino médio e que residem nas regiões metropolitanas de Belo Horizonte, São Paulo, Rio de Janeiro, Salvador, Recife e Porto Alegre foram discutidos à luz dos resultados obtidos.

No sexto capítulo, a nota do ENEM 2008 foi imputada para os indivíduos que entraram na base de dados da PME entre 2008 e 2009, utilizando a metodologia de Elbers et al (2003) e adaptada por Ribas e Machado (2009). Os resultados da foram apresentados, discutidos e, posteriormente, utilizados para estimar o modelo probabilístico multinomial logit, para a transição da situação ocupacional dos jovens. Com isso, podemos discutir a influência da proficiência cognitiva dos mesmos sobre a sua situação no mercado de trabalho e sobre a continuação dos estudos no ensino superior.

Por fim, temos a conclusão no sétimo capítulo, onde os principais resultados foram sumarizados e as considerações finais foram feitas.

2 REVISÃO BIBLIOGRÁFICA

Esse capítulo busca fazer uma revisão bibliográfica dos trabalhos que contribuíram para o entendimento e avanço da economia da educação e está dividido em duas partes. A primeira discute os pontos principais da Teoria do Capital Humano, bem como as críticas feitas a ela. A segunda parte trata dos fatores associados à proficiência escolar dos estudantes, com base na habilidade cognitiva e não-cognitiva dos mesmos, tendo-se por referência a função de produção educacional.

2.1 Sobre a Teoria do Capital Humano e suas Críticas

O papel que alguns atributos pessoais podem exercer sobre as escolhas dos indivíduos, bem como a determinação de sua renda, é um tema muito discutido e fundamentado na literatura econômica. Em especial, podemos destacar o papel que a educação formal possui sobre as decisões e oportunidades pessoais ao longo da vida.

O modelo elaborado por Mincer (1958) introduziu a ideia de que a diferença de renda entre trabalhadores é determinada por três fatores: a quantidade de treinamento e a sua experiência profissional. A experiência seria função da idade e do tempo de profissão, mas possui rendimentos marginais decrescentes devidos, principalmente, ao avanço tecnológico e também ao desgaste natural biológico que afeta os seres humanos. Logo, o autor destaca que existe uma diferença de renda entre ocupações que demandam níveis educacionais diferenciados e, ainda, uma diferença de renda dentro de uma mesma ocupação, devido à experiência e idade.

Alguns anos depois, Becker (1962) sugeriu a importância da escolaridade para explicar a desigualdade de renda entre indivíduos. Essa ideia foi inovadora em relação a outras de sua época porque, geralmente, a riqueza pessoal era

vinculada à posse de capital físico, e assim, essa era considerada a principal fonte de desigualdade de renda entre as famílias.

A principal contribuição de Becker para o estudo do capital humano foi a elaboração do conceito da taxa de retorno da educação sobre o rendimento dos indivíduos. Se o investimento presente em educação reverte-se em renda futura, então a taxa de retorno é uma medida de custo-benefício que o investimento monetário e custos de oportunidades durante o período escolar propiciam após a inserção do indivíduo no mercado de trabalho.

A discussão teórica da contribuição do conhecimento na determinação de salários é bastante rica. De um lado, temos a abordagem do capital humano que apresenta a saúde, habilidade, experiência e escolaridade como pontos centrais para explicar a desigualdade de renda entre os indivíduos com níveis educacionais diferentes. No entanto, existem trabalhos (BOWLES et. al, 2001; HECKMAN & RUBINSTEIN, 2001) que identificam o problema da educação e da distribuição de salários no mercado de trabalho também por meio de outros mecanismos, chamados de habilidades não-cognitivas, que diferenciam indivíduos no mercado de trabalho.

A ideia desses autores é que, como os indivíduos passam boa parte de sua juventude na escola, pode haver o desenvolvimento de outras habilidades individuais, que dependem de sua presença na turma. A vivência com pessoas da mesma coorte, provenientes de classes sociais diferentes e com alunos mais habilidosos, forma a rede social de amigos e influência mútua que, posteriormente, pode influenciar a inserção no mercado de trabalho muito mais do que as habilidades cognitivas desenvolvidas na escola.

Para Heckman & Rubinstein (2001), as pessoas com alto nível de QI (Quociente de Inteligência) podem não alcançar sucesso em vida por não possuir disciplina ou concentração suficiente para resolver problemas. Nesse sentido, um alto nível de escolaridade pode sinalizar que os indivíduos são mais perseverantes que os demais. Para os autores, se a formação das habilidades cognitivas das pessoas se completa até uma idade precoce, seria demasiadamente ineficaz investir somente na dimensão cognitiva ao longo de todo ciclo escolar, enquanto que o

custo de aprendizado de habilidades não-cognitivas em idades mais avançadas é menor.

A principal crítica à Teoria do Capital Humano, no entanto, baseia-se na Teoria da Sinalização no mercado de trabalho, de Spence (1973). O seu principal interesse recai sobre as relações de trabalho pela óptica da assimetria de informação entre trabalhadores e firmas atuando no mercado de trabalho.

O autor explica que o empregador não conhece de fato as reais características, que são pertinentes ao trabalho, do candidato à vaga de emprego e nem como ele se comportará após sua possível admissão. O candidato, por sua vez, possui traços pessoais imutáveis, como sexo, raça e, principalmente, habilidade cognitiva inata, bem como características sob seu controle, como a educação, que são observáveis pelo empregador durante uma entrevista.

O fato de alguns candidatos possuírem mais escolaridade que outros, para Spence (1973), sinaliza que estes possuem mais habilidade cognitiva, sendo este um atributo pessoal constante ao longo do tempo, correlacionado positivamente com a escolaridade. A habilidade cognitiva garante um menor custo de concluir ciclos de educação formal. Assim, o autor considera que a escolaridade é a melhor credencial sobre a aptidão do candidato, apesar de outros traços serem importantes, mas não observáveis até o momento da contratação.

O chamado *sheepskin effect*, ou efeito diploma, derivado da teoria da sinalização, significa que a renda do indivíduo aumenta segundo sua escolaridade por discriminar a habilidade, mas aumenta mais quando esse ano adicional coincide com a obtenção de um diploma de conclusão. Isso significa uma relação não linear e descontínua entre educação e a renda (CRESPO & REIS, 2006, pág.25). Para o caso brasileiro, Crespo e Reis (2009) encontram um efeito diploma positivo, ainda que demonstrem que o efeito diminuiu no período analisado (1982 a 2004), conforme o aumento da oferta de educação no país.

A afirmação de que a única função da educação é diferenciar pessoas com habilidades (inatas) que são valorizadas pelo mercado de trabalho é criticada, pois, caso seja verdadeira, poderia existir um mecanismo mais barato do que o sistema educacional para identificar trabalhadores habilitados. Assim, não

podemos afirmar que os indivíduos nascem com certo nível de habilidade e passam pelo sistema educacional sem serem afetados pelo mesmo, tampouco podemos afirmar que o sistema de ensino unicamente aprimora as habilidades cognitivas que são valorizadas no mercado de trabalho (BLAUG, 1992; LAYARD & PASCHAROPOULOS, 1974).

2.2 Fatores Associados ao Desempenho Escolar

A proficiência escolar possui uma correlação positiva com o salário do trabalhador, e possui vários determinantes, bem como pode funcionar como motivadora de uma série de outras escolhas dos indivíduos. A escolha ocupacional, a residência em locais diferenciados, os padrões de consumo e a escolha dos amigos são exemplos de tais escolhas (GHEZ, 1975). Como a educação possui um papel tão vital para as pessoas, devemos investigar os determinantes para se investir tempo e recursos monetários nesse processo, para entender porque as pessoas simplesmente não optam por completar o ciclo de educação formal.

Segundo Hanushek (2010), a medida mais comumente usada como produto educacional é a escolaridade do indivíduo, isto é, a quantidade de anos de estudos completos. Contudo, essa medida simplesmente contabiliza os anos cursados, sem considerar a qualidade do ensino. Assim, uma métrica mais adequada seria a utilização de resultados individuais de exames padronizados de proficiência para explicar o conhecimento adquirido na escola.

Os fatores que afetam a proficiência dos alunos em sua vida acadêmica são muitos e, por vezes, correlacionados entre si. Como exemplo, citamos a relação entre um produto educacional, explicada por um vetor de variáveis pessoais (sexo, gênero, cor, etc.), familiares (nível socioeconômico familiar, arranjo familiar, número de moradores no mesmo domicílio) e insumos escolares (escolaridade dos professores, infraestrutura da escola, etc.).

A função de produção educacional que explica a escolaridade “**E**” do estudante com por meio de fatores cognitivos “**C**”, como a proficiência ou escolaridade, e não-cognitivos “**N**”, como gênero, cor/raça, idade, background familiar e insumos

escolares, somado a um erro aleatório “v”, como explica Soares (2010), pode ser escrita da seguinte maneira:

$$E = \gamma_n N + \gamma_c C + v \quad (2.1)$$

A escolaridade em si pode não ser o objetivo da existência do sistema educacional. O foco das políticas de ensino seria produzir capital humano, isto é, aumentar a proficiência cognitiva dos alunos a cada ano adicional de escolaridade. Assim, muitos autores utilizam a função de produção educacional como uma forma de expressar o desempenho escolar, e não a escolaridade, em função de fatores cognitivos e não-cognitivos, como explica Albernaz et al. (2002):

$$y = F(c, m, g, p, s) \quad (2.2)$$

Nesse caso, “y” significa o desempenho do aluno, medido em alguma escala de proficiência, “c” representa características pessoais, “m” os fatores familiares, “g” a influência dos colegas de classe, “p” as variáveis ligadas aos professores e “s” as características da escola que o aluno frequenta.

No caso brasileiro, existem diversos trabalhos, como Barbosa e Fernandes (2001), César e Soares (2001), Albernaz et al. (2002), Soares (2004) e Franco et al. (2007) que utilizam a nota obtida no Saeb (Sistema de Avaliação da Educação Básica) para alunos da 4ª e 8ª série¹ do ensino fundamental, para estimar o desempenho escolar do aluno em função de variáveis pessoais e da escola. No caso de Minas Gerais, Machado et al. (2005) e Soares (2005) utilizam uma base de avaliação específica do Estado, o SIMAVE (Sistema Mineiro de Avaliação da Educação Pública), para o mesmo objetivo.

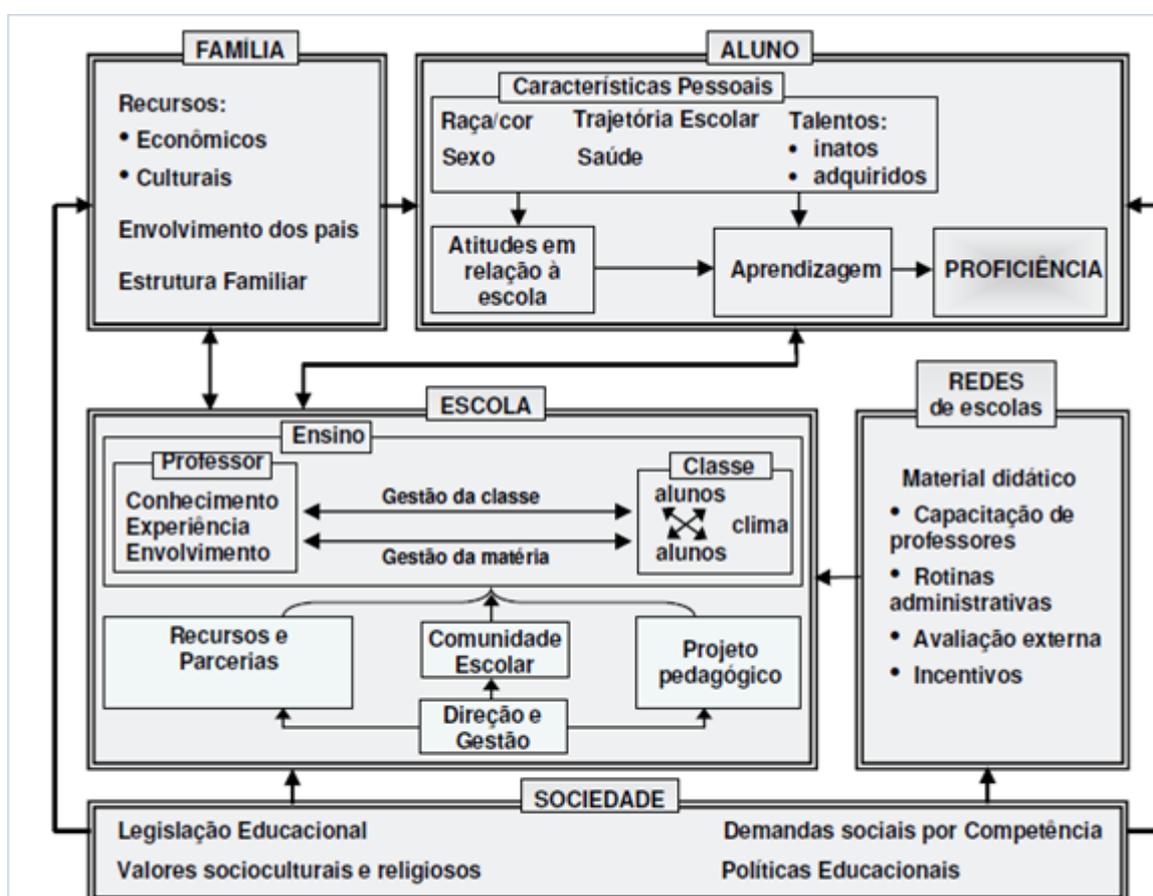
Para estimar os fatores associados à proficiência dos alunos, Golgher (2010) utiliza a base de dados do ENEM, agregando diversos atributos pessoais por meio de análise multivariada e estuda sua influência na proficiência dos alunos no exame, por meio do método de MQO e de Heckman. O seu resultado mostra que

¹ Equivalentes, respectivamente, ao 5º e 9º ano do ensino fundamental, segundo o novo formato do ensino fundamental.

o viés de seleção é menor em escolas privadas e federais, e os fatores não-observáveis que estão associados com a maior probabilidade de presença do aluno no exame também estão associados com maiores notas na prova.

No contexto da realidade brasileira, queremos investigar quais são os determinantes do desempenho dos alunos, isto é, estimar a função de produção educacional. Dessa forma, a Figura 1 apresenta o fluxo de relações de influência entre fatores de diferentes instituições que podem determinar a proficiência do estudante e que são discutidos em separado a seguir:

Figura 1 - Modelo conceitual explicativo da proficiência cognitiva



Fonte: Andrade (2008) apud Soares (2008).

A legislação educacional, bem como os costumes e valores da sociedade brasileira influenciam, possivelmente, a forma de funcionamento institucional das escolas, a satisfação dos pais de estudantes e estes próprios sobre o que se aprende na escola e o valor desse conhecimento. Embora sua importância seja fundamental, o papel das variáveis relativas à sociedade, de pouco controle dos

estudantes e de seus pais, foge ao escopo deste trabalho. Assim, buscamos concentrar nas dimensões mais próximas da realidade dos jovens, como suas próprias características, a família e a escola que ele frequenta.

2.2.1 Características Pessoais

As características pessoais dos estudantes, tais como sexo, raça, saúde e habilidades inatas podem influenciar em seu resultado educacional (proficiência) de diversas maneiras. Muitos trabalhos baseados no Saeb (ANDRADE R.J., 2008; ALBERNAZ et al., 2002; CÉSAR & SOARES, 2001; FRANCO et al., 2007) apresentam uma correlação significativa entre desempenho escolar e gênero, mas ainda que não sejam unânimes com respeito aos alunos da 4ª série, apontam que os alunos do sexo masculino da 8ª série de ensino fundamental e 3º ano do ensino médio possuem uma vantagem em relação ao sexo oposto quando testados sobre conhecimentos de matemática.

Um outro exemplo da diferença de desempenho entre sexos pode ser identificado no PISA (Programme for International Student Assessment) realizado, trienalmente, pela OCDE (Organização para Cooperação e Desenvolvimento Econômico). Considerando a edição de 2006, foram testados mais de quatrocentos mil jovens de 57 países diferentes (PISA, 2006). O resultado do exame de leitura do PISA 2006 apontou um desempenho melhor para o sexo feminino em todos os países participantes do exame. No entanto, na prova de matemática, observa-se que, em 31 dos 57 países participantes, o sexo masculino apresentou um desempenho superior ao feminino, e em 21 outros países não houve diferença. No Brasil, o sexo feminino apresentou um desempenho melhor na prova de leitura e compreensão do PISA 2006, mas, na prova de matemática, ocorreu o contrário (PISA, 2006).

Com respeito à desigualdade de proficiência e oportunidades de estudo de pretos e pardos, em relação a brancos e amarelos, a maioria dos trabalhos inclui essa variável como determinante do desempenho do estudante, mesmo que haja um erro de medida por auto-atribuição, uma vez que o candidato é livre para responder o que quiser, segundo Andrade (2008). As causas pelas quais a raça

interfere na proficiência do aluno podem refletir as desigualdades sociais as quais pretos e pardos enfrentam, desde oportunidades de trabalho, nível socioeconômico familiar ou residência em locais de pior acesso a escolas de qualidade.

O resultado encontrado sistematicamente para diversas edições do exame do Saeb foi uma correlação negativa entre a declaração do aluno ser preto ou pardo e a sua nota no exame, em relação a brancos e amarelos, para qualquer série cursada, tanto em escolas públicas quanto em escolas privadas (ANDRADE, 2008; ALBERNAZ et al., 2002; CÉSAR & SOARES, 2001; FERRÃO et al, 2002; GONÇALVES & FRANÇA, 2008; FERREIRA, 2008; FRANCO, 2009). De acordo com os resultados obtidos por ALBERNAZ et al. (2002), isto pode indicar que a desigualdade de proficiência entre brancos e não-brancos não é decorrida apenas da desigualdade de oportunidades, mas podem haver outros fatores.

Ao aprofundar a investigação do efeito da cor/raça sobre a vida escolar dos indivíduos, Guimarães & Rios-Neto (2010) investigam, no Brasil, a hipótese de que as origens sociais determinam a probabilidade de transitar entre séries escolares. Uma consequência de sua afirmação seria um efeito causal declinante da raça sobre a transição escolar ao longo dos anos, devido à seleção a cada fim de ano letivo.

A confirmação dessa hipótese teria como resultado, ao final do ciclo básico de ensino, a presença de uma certa homogeneidade das características pessoais dos estudantes no último ano. No entanto, os autores afirmam que não é possível postular que essa hipótese seja correta, isto é, o efeito não é declinante, apesar de outros estudos citados pelos autores encontrarem um resultado contrário.

A respeito dos estudos similares obtidos por meio do exame do ENEM de 2005, Golgher (2010) apresenta resultados que corroboram o fato de que não-brancos possuem menor rendimento escolar que os brancos, ressaltando que estes formam a maioria nas redes de ensino privada e federal e, aqueles nas redes estadual e municipal.

A comparação da transmissão intergeracional de educação entre negros e brancos comprova a maior persistência entre negros. Em outras palavras, a

probabilidade de uma pessoa cujo pai não frequentou a escola também não frequentar, é maior entre negros do que entre brancos. Se considerarmos as diferenças regionais brasileiras, a maior persistência observada no Nordeste em relação ao Sudeste está fortemente associada à maior probabilidade, na primeira região, de um pai sem escolaridade transmitir essa condição a seus filhos (FERREIRA & VELOSO, 2003).

A hipótese da transmissão intergeracional da educação postula, então, que a educação dos filhos é determinada pela educação dos pais. Além de influenciar diretamente na nota do exame de proficiência, este atributo também afeta o desempenho escolar dos filhos de outras maneiras. De acordo com Hanushek (2008), os estudantes cujos pais possuem relativamente mais escolaridade que outros tendem a prolongar sua formação acadêmica e ascender de nível socioeconômico porque o nível educacional dos pais é positivamente correlacionado com a renda permanente da família.

A mobilidade intergeracional de educação de diferentes países pode ser observada no quadro 1, onde Ferreira & Veloso (2003) resumem os resultados de diversos outros trabalhos que estimam o grau de persistência educacional, os coeficientes de MQO da escolaridade dos pais como variáveis explicativas para a escolaridade dos filhos. A principal conclusão é que tal persistência é maior na América Latina e, no Brasil, é uma característica bastante marcante.

QUADRO 1 – Estudos de persistência intergeracional de educação

Autor	Grau de Persistência Educacional	País
Borjas (1992)	0,25	Estados Unidos
Cough e Dunn (1997)	0,27	Estados Unidos
Mulligan (1997)	0,32	Estados Unidos
Behrman, Gaviria e Székely (2001)	0,35	Estados Unidos
Cough e Dunn (1997)	0,20	Alemanha
Behrman, Gaviria e Székely (2001)	0,70	Brasil
Behrman, Gaviria e Székely (2001)	0,70	Colômbia
Behrman, Gaviria e Székely (2001)	0,50	México
Behrman, Gaviria e Székely (2001)	0,50	Peru
Lilliard e Willis (1994)	0,19	Malásia

Fonte: Ferreira & Veloso (2003, pág. 484)

Muitos trabalhos comprovam empiricamente a influência positiva da educação dos pais como determinante da educação dos filhos. Esse é o caso de Marteleto (2002), utilizando dados da PNAD de 1977 a 1997 para estudar a influência da escolaridade da mãe em diferentes coortes, bem como os trabalhos que estimam a influência direta deste fator sobre a proficiência do aluno, por meio de funções de produção educacional, como Riani e Rios-Neto (2008), Menezes-Filho (2007) e Machado et al. (2008).

Em suma, como explica Ferreira & Veloso (2003), a persistência da educação dos pais sobre os filhos existe no Brasil, e é bastante forte nos extremos da distribuição de escolaridade. Isto é, pais com pouca ou nenhuma instrução transmitem sua posição aos seus filhos mais fortemente do que aqueles que possuem uma escolaridade média. Por outro lado, para os pais que se encontram no topo da distribuição de escolaridade, que completaram nível muito elevado no sistema educacional, a probabilidade de que seus filhos alcancem tal nível é maior que a média.

2.2.2 Nível Socioeconômico

No que se refere aos recursos familiares à disposição de seus membros mais jovens, Silva e Hasenbalg (2002) distinguem pelo menos dois canais de influência da educação dos pais sobre a educação de seus filhos. O primeiro seria a renda familiar e condições de moradia, que podem propiciar ao jovem um bom ambiente de estudos, acesso ao ensino de qualidade. O segundo seria a valorização da educação dos filhos como um bem final, devido ao maior capital humano na família, o que poderia aumentar a sua demanda educacional.

Ao considerar a importância do capital econômico familiar sobre o aprendizado e capacitação de seus membros, devemos considerar a existência de uma demanda por um mercado de crédito, onde as pessoas podem tomar recursos para financiar sua educação, sob o ponto de vista que este é um investimento em capital humano. Desse modo, a eficiência do mercado de crédito também direciona as decisões intrafamiliares. Para Assunção e Alves (2007), a decisão dos pais em investir em educação dos filhos depende dessa eficiência, e se

baseia na apreciação do valor presente líquido do custo de estudar e da renda futura. Se o mercado de capitais operar ineficientemente, o investimento passa a depender mais fortemente da riqueza familiar disponível.

O trabalho dos autores discute uma vasta literatura sobre restrições de crédito e trabalho infantil, a qual afirma que o trabalho infantil pode ser usado para compensar choques econômicos adversos dentro da família, ou seja, o fato de um dos pais ficar desempregado pode aumentar a probabilidade dos filhos deixarem a escola e entrarem no mercado de trabalho. Além disso, utilizando a PNAD (Pesquisa Nacional por Amostra de Domicílios) de 1981 a 2001 para estudar a relação entre restrição de crédito e investimento em educação intrafamiliar, eles concluem que a qualidade da educação infantil e a eficiência do mercado de crédito se correlacionam na mesma direção.

2.2.3 Estrutura e Arranjo Familiar

A família possui uma grande influência sobre as decisões de seus membros e, por isso, talvez ela seja um dos fatores de maior peso sobre o desempenho dos indivíduos na escola. Para Astone e McInahan (1994), Silva & Hasenbalg (2002) e Heckman & Masterov (2005), os laços familiares importam para o aproveitamento escolar de seus membros e uma ruptura no arranjo familiar típico, em que a criança mora com ambos os pais, pode influenciar negativamente o desempenho do mesmo na escola, sendo o efeito potencializado caso a família esteja exposta aos fatores de restrições orçamentárias.

Dentro da estrutura de um domicílio, as pessoas podem ter diferentes posições familiares. Geralmente, o estudante vive em um domicílio juntamente com os pais e/ou outros parentes. Além disso, ele pode dividir o domicílio com irmãos, que podem influenciar a alocação de recursos educacionais dentro da família. Isso poderia resultar em um estoque de capital humano disponível para cada jovem de forma diferenciada dentro da mesma residência, ou de forma igualitária, mas em menor nível, acarretando em uma correlação negativa entre o número de irmãos que um indivíduo possui e seu desempenho escolar (HANUSHEK, 1992; PATRINOS & PASHAROPOULOS, 1997).

Para Patrinos e Pasharopoulos (1997), a relação negativa entre tamanho da família e rendimento escolar, identificado em países em desenvolvimento, é fruto do mecanismo de diluição de recursos entre irmãos. Isso pode ser um fator determinante da escolha alocativa dos pais em relação ao trabalho e estudo dos filhos e, assim, eles incentivam alguns dos filhos a trabalhar, enquanto outros somente a estudar. Nesse caso, a família maximiza a utilidade total do grupo, o que significa uma alocação maior de recursos para aqueles filhos que possuem um maior retorno futuro sobre a educação, causando um subinvestimento nos menos aptos.

Este caso, o do investimento desigual entre filhos, seria o mais comum no Brasil, segundo Patrinos & Pasharopoulos (1997) e Marteleto (2002), ainda que não exista um padrão bem definido para esta relação, considerando outros países em desenvolvimento. Apesar disso, Hanushek (1992) aponta outras duas alternativas de investimento intrafamiliar. A família pode não discriminar a alocação de tempo entre os filhos e, então, uma quantidade maior de filhos dilui o tempo gasto com cada um individualmente, mantendo constante o tempo disponível para cuidar dos filhos.

A outra situação, de acordo com o autor, ocorre se os pais fizerem uma alocação compensatória do tempo gasto individualmente entre os filhos, movida por forças contrárias. Uma dessas forças é a maximização do desempenho médio dos filhos, que proporciona maior investimento nos mais aptos. A outra força é a minimização da variância, que transfere parte do tempo disponibilizado dos mais aos menos aptos. No entanto, em qualquer situação, o ajuste compensatório é menor quanto maior for a família, devido justamente ao estoque limitado de tempo destinado à educação dos filhos.

O autor também destaca a influência da ordem de nascimento dos filhos dentro da família como determinantes de seu desempenho escolar. Em famílias grandes, o desempenho dos filhos é menor, quanto mais irmãos mais novos eles possuírem, e é menor quanto mais próxima for a idade entre eles, devido à disputa dos recursos educacionais.

2.3 Determinantes Escolares da Proficiência Cognitiva

As instituições de ensino que compõem o sistema educacional brasileiro têm por objetivo incentivar a geração de habilidades cognitivas na forma de conhecimento técnico ou acadêmico-científico, sendo este último o conteúdo prevalecente. Conforme Goldin (2001), outros países, como a Alemanha, oferecem modelidades majoritariamente técnicas de ensino médio, ou os Estados Unidos, que mistura o ensino acadêmico, genérico, com disciplinas práticas ao mercado de trabalho.

As escolas brasileiras utilizam-se de sua infraestrutura e docentes, cujas qualidades variam, que podem influenciar todos os seus alunos. Uma das primeiras obras da literatura sobre qualidade da educação a inspecionar tais influências foi o relatório Coleman et al. (1966), que apontou como o sistema educacional norte-americano produzia desigualdades entre estudantes de níveis sócioeconômicos diferentes dentro da mesma escola, bem como a existência de uma possível relação entre a heterogeneidade dos alunos e seu rendimento escolar.

Para estudar a importância desses determinantes no Brasil, Albernaz et al. (2002) estimaram uma função de produção educacional da proficiência do indivíduo em função de características pessoais e escolares por meio de modelos lineares hierárquicos. O principal resultado encontrado foi que o melhor desempenho no Saeb 1999 pertencia a escolas que concentravam alunos com nível socioeconômico acima da média, o que reflete a existência de uma auto-seleção nas escolas.

O nível socioeconômico também influencia o aproveitamento do aluno na escola, e mais uma vez, isso se reflete em sua proficiência. A influência da escolaridade do professor sobre a proficiência do aluno é crescente em relação ao nível socioeconômico do aluno. Além disso, controlando por todos esses efeitos, as escolas privadas obtiveram vantagem em relação às públicas. As variáveis de controle da escola, como infraestrutura, características do diretor e do professor possuem pouca influência sobre a proficiência dos alunos, apesar do salário dos docentes ser importante na rede privada e o número de horas-aula ter um efeito positivo sobre o desempenho dos alunos (ALBERNAZ et al., 2002).

Os modelos hierárquicos estimados por Cesar e Soares (2001) para o Saeb 1999 reforçam essa relação. As escolas concentram alunos de nível sócioeconômico parecidos, mais ainda, quanto maior o desvio no nível social de um aluno em relação à média de sua escola, maior é o seu aproveitamento, o que pode ser caracterizado como um efeito gerador de desigualdades de aprendizado dentro da escola. Os resultados do autor quanto ao efeito positivo da heterogeneidade na sala de aula também são pouco otimistas: “o aluno pobre na escola rica é beneficiado pelo meio. Entretanto, o ganho do aluno pobre é muito inferior à perda do aluno rico” (CESAR & SOARES, 2001, pág.108).

Ao tentar dividir a variação devida a fatores escolares e pessoais, Menezes-Filho (2007) mostra que a variância da nota do Saeb 1999 que é explicada por fatores escolares difere entre Estados. Assim, existe uma diferença regional para a qualidade das escolas medidas em termos da nota do estudante, e aponta que as escolas do Distrito Federal são as mais homogêneas do país, enquanto que as escolas do Tocantins possuem um grau de heterogeneidade bastante elevado.

O autor estima um modelo somente com variáveis da escola e outro com variáveis pessoais e escolares e afirma que, ao incluir as características pessoais e familiares em seu exercício econométrico, a parcela da variação explicada por atributos escolares decresce. Isso decorre do fato de que algumas escolas recebem alunos cujo nível socioeconômico está acima da média populacional e, por isso, a diferença da nota média de uma escola pode refletir, na verdade, o nível socioeconômico médio daquela instituição.

No que diz respeito às diferenças de escolas públicas de Estados diferentes, não é possível afirmar que a quantidade de recursos financeiros destinados a políticas educacionais esteja relacionada com o desempenho dos alunos das escolas dessa região. No entanto, a gestão pública dos recursos destinados à educação pode ter mais peso sobre a eficiência do ensino público estadual do que a magnitude dos recursos financeiros (MENEZES-FILHO,2007).

2.4 Determinantes Ocupacionais pós-Ensino Médio

As aspirações educacionais e ocupacionais dos jovens são temas atuais e recorrentes no debate econômico (Hardie, 2009; Schoon & Polek, 2011; Collier et al, 2011). Os indivíduos que possuem metas mais ambiciosas para sua própria educação e ocupação acabam por conseguir um nível educacional e ocupacional elevado, quando comparados à média dos jovens, assim como a elevada proficiência escolar aumenta a probabilidade do estudante optar por uma trajetória que envolva mais estudos após o término do ciclo básico de educação.

Na literatura internacional, Collier et al. (2011) investigam a relação entre o desempenho de alunos da Inglaterra e país de Gales em exames de proficiência prestados ao término do ciclo obrigatório de ensino e sua escolha sobre o tipo de educação após essa etapa. Os estudantes podem escolher entre o ensino profissionalizante e o acadêmico, que é exigido para o ingresso no ensino superior. Para este objetivo, os autores utilizaram o método de variáveis instrumentais para controlar o problema de endogeneidade entre a escolha educacional e a proficiência.

Os autores observaram que a proficiência no exame feito ao término do ciclo obrigatório é bastante correlacionada com a escolha acadêmica, mas o *background* familiar possui pouco poder explicativo. Isso aconteceria porque os pais influenciam a escolaridade e a proficiência dos filhos durante as etapas mais precoces da sua vida escolar e não diretamente a proficiência no exame e/ou a escolha no mercado de trabalho.

O sucesso nos exames de proficiência pode direcionar as decisões sobre o futuro dos jovens, mas tais decisões podem também influenciar o resultado nos exames. Se a escolha futura exigir um alto nível de capital humano, de tal forma que o jovem se prepara com antecedência, então tal preparação é captada por tais exames. Se as escolhas dependem de um processo seletivo, para o qual o exame de proficiência é utilizado, então as decisões sobre o futuro influenciam no resultado do exame.

A relação entre o desempenho em exames de proficiência e a decisão do aluno em continuar seus estudos, cria um viés de simultaneidade. Tanto avançar no ciclo de educação formal, quanto o desejo de cursar o ensino superior afetam o resultado obtido nos exames (COLLIER et al., 2011, pág.16).

O trabalho de Hardie (2009) utiliza dados de painel representativos de jovens de 14 a 22 anos, de 1979 a 1994, para os Estados Unidos, com o objetivo de estimar a influência do background familiar, eventos críticos (casamento, filhos, divórcio, doenças e prisão) e outras variáveis de controle sobre a falha em conseguir alcançar seus planos educacionais e ocupacionais. O resultado evidencia que a situação ocupacional dos pais acarreta efeitos positivos sobre a ocupação dos filhos. No entanto, a escolaridade dos pais possui pouca correlação com a assertividade das suas escolhas ocupacionais, exceto quando eles possuem baixa escolaridade, caso em que a assertividade dos objetivos dos filhos é bastante reduzida. Por fim, a proficiência educacional dos alunos é positivamente correlacionada com a propensão a atingirem suas metas ocupacionais.

Por meio de dois painéis representativos para o Reino Unido, Schoon & Polek (2011) afirmam que as aspirações são determinadas pela sua inserção na escola e no trabalho, bem como o nível socioeconômico, capacidade cognitiva e proficiência escolar. As expectativas dos adolescentes são importantes previsores de seu *status* futuro, e assim, os mais ambiciosos são mais propensos a conseguirem uma posição diferenciada no mercado de trabalho. Os resultados encontrados mostram que, para as coortes mais jovens, o papel da capacidade cognitiva e do *status* social familiar possui maior peso na formação dessa ambição.

A educação promove a elevação da renda dos indivíduos, mas de forma lenta, se considerarmos todo o ciclo pelo qual as pessoas passam no ensino fundamental, médio e superior. O trabalho de Barros *et al.* (1997) investiga por que pessoas com grau de escolaridade baixo conseguem ultrapassar a linha da pobreza no Brasil, e afirma que existem fatores que contribuem para formar uma heterogeneidade entre trabalhadores dentro do subgrupo com baixa escolaridade que é similar à heterogeneidade de renda encontrada na população. Para entender como funciona a heterogeneidade de ocupações entre pessoas com o

mesmo grau de instrução, os autores citam três situações que explicariam essas diferenças.

Na primeira delas, pessoas com as mesmas oportunidades de acesso às diferentes ocupações possuem rendas diferenciadas. Isso seria o caso em que cada pessoa escolhe a ocupação que mais lhe satisfaz e é remunerado de acordo com as características daquela função. Na segunda situação, ocorreria heterogeneidade por meio de discriminação ou acesso privilegiado ao mercado de trabalho, limitando as opções de escolha de certos grupos de pessoas.

Por fim, podemos considerar uma situação em que os trabalhadores possuem produtividades diferenciadas, e isso influenciaria o acesso a certas ocupações para um subgrupo composto por pessoas que detêm certos atributos produtivos necessários para a execução da função.

Ao estudar a escolha de alocação do tempo do jovem entre educação, trabalho e lazer, devemos considerar que algumas características pessoais podem influenciar esta decisão. Neste aspecto, os fatores cognitivos e não-cognitivos podem determinar o custo de alocar o tempo em determinada atividade. Segundo Ordine e Rose (2011), se a utilidade de um indivíduo é determinada pela alocação do tempo entre escola e trabalho, devemos considerar o custo desse indivíduo trabalhar e estudar, bem como as habilidades individuais (cognitivas e não-cognitivas) que diminuem esse custo.

No que diz respeito à relação entre proficiência escolar e mercado de trabalho, Curi & Menezes Filho (2006) constroem um pseudo-painel para estudar a proficiência média da coorte nascida entre 1977 e 1978 que fizeram a prova do SAEB em 1995 e os salários médios da coorte cinco anos depois, utilizando uma imputação da nota estimada do SAEB no CENSO 2000. O resultado encontrado foi que a proficiência média da coorte, bem como insumos escolares, são positivamente correlacionados com o salário daquela coorte cinco anos após deixarem o ensino médio.

Para Soares (2010), que utilizou um método de imputação da nota do ENEM (2003 a 2007) na PME (2003-2008) e estimou um modelo probabilístico para o comportamento dos jovens ao longo de doze meses, a proficiência cognitiva é

eficaz em aumentar a probabilidade do jovem trabalhar, mas não dele continuar estudando. Além disso, os retornos econômicos da educação, por via da proficiência, são positivos, mas baixos, quando comparados com outros países em desenvolvimento.

Segundo Eckstein e Wolpin (1999), o mercado de trabalho oferece salários que podem ou não ser correlacionados com o desempenho escolar. Segundo a análise dos autores, os jovens que deixam a escola são menos motivados e possuem uma menor habilidade escolar, justamente aquela que diminui o custo de frequentar a escola e, por isso, são mais propensos em entrar precocemente no mercado de trabalho.

O fluxo de entrada e saída de jovens no mercado de trabalho pode ser elevado, como afirmam Clark e Summers (1982), e isso pode ser mais intenso entre os jovens que estão estudando, uma vez que podem deixar de trabalhar e continuar na situação de estudante, ao invés de desempregado.

Considerando toda discussão sobre educação anterior, a proposta desta dissertação é investigar empiricamente os determinantes da proficiência cognitiva utilizando a base de dados do ENEM 2008, cujo uso para este fim é recente. Além disso, estudamos como a proficiência cognitiva influencia a situação ocupacional após a conclusão do ciclo básico de educação, como investigado por Soares (2010).

3 METODOLOGIA

A função de produção educacional relaciona a proficiência com fatores pessoais, familiares e escolares. Matematicamente, ela possui caráter aditivo em relação ao efeito de cada variável, o que nos permite identificar o impacto parcial de cada fator sobre a proficiência do jovem.

Ao observar os dados declarados pelos jovens no questionário do ENEM, temos que considerar a existência de um agrupamento não-aleatório de alunos dentro da mesma escola, que pode influenciar todos os seus alunos de uma maneira homogênea. Outro ponto relevante em nossa análise é o aspecto auto-seletivo da prova do ENEM, já que os estudantes não são obrigados a prestar o exame, o que pode ser uma fonte de viés, caso essa ausência não seja aleatória entre os indivíduos.

O nosso objetivo de estimar a função de produção educacional foi alcançado pelo método dos mínimos quadrados ordinários (MQO), a princípio, para uma análise preliminar dos dados, desconsiderando as fontes de viés citadas. A extensão natural da nossa análise foi introduzir métodos que permitissem corrigir o viés de auto-seleção e de agrupamento não-aleatório nas escolas.

Para o tratamento do possível viés de auto-seleção, utilizou-se o método de correção de Heckman em dois estágios. Este método consiste em estimar, em primeiro lugar, a probabilidade de o candidato comparecer ao exame, por meio de um modelo Probit, em função de variáveis explicativas cuja variável dependente é binária e referente à presença na prova. Em seguida, estima-se uma regressão linear da nota do ENEM em função de uma série de variáveis explicativas, em que o viés é corrigido por meio da introdução da razão inversa de Mills, calculada com base no primeiro estágio, como uma variável do modelo.

Por fim, considerando a quebra de uma das hipóteses do MQO, a de que as amostras são independentes e identicamente distribuídas, pelo agrupamento de jovens com características semelhantes na mesma instituição de ensino, foram estimados modelos de regressão hierárquico linear (HLM).

A lógica desses modelos é a seguinte: se fosse estimado um modelo MQO para cada escola, teríamos um coeficiente que representaria a nota média daquela escola, mas como não estamos interessados em cada escola em particular, os modelos HLM fornecem uma maneira sintética de atribuir um efeito aleatório à nota média, que se refere à diferença de cada escola, e também permitem explorar o uso de variáveis explicativas do nível escolar sobre os seus alunos².

A etapa seguinte refere-se à imputação da nota da prova objetiva do ENEM 2008 aos indivíduos que estão cursando o último ano do ensino médio na PME de 2008 e 2009. Esse objetivo foi alcançado por meio da metodologia a ser discutida, permitindo, com isso, estudar a relação entre a proficiência cognitiva e a posição no mercado de trabalho.

A princípio, são identificadas variáveis comuns a ambas as bases de dados. A nota da prova objetiva do ENEM é estimada por um subgrupo das variáveis utilizadas na análise anterior. Em seguida, os coeficientes foram utilizados para estimar a proficiência dos indivíduos na base da PME. Assim, consideramos que não há, pelo menos no prazo de um ano, diferenças significativas, entre 2008 e 2009, sobre o efeito das variáveis utilizadas na imputação, uma vez que a proficiência estimada para os indivíduos em 2009 é feita com base nas informações do ENEM 2008.

Por fim a proficiência cognitiva estimada entra como uma variável explicativa do modelo logit multinomial, que foi usado para estimar a probabilidade de transição de estudantes do ensino médio entre diferentes formas de inserção no mercado de trabalho e para a continuidade da educação.

O trabalho de Soares (2010) utiliza outra metodologia para o mesmo fim. O autor afirma que é possível fazer uma identificação quase exata dos indivíduos nas duas bases, por meio das variáveis de região metropolitana, data de nascimento, sexo e número de pessoas no domicílio. No entanto, esse método não garante a identificação exata e o número de observações se reduz consideravelmente.

² Tentou-se também utilizar a razão inversa de Mills no contexto dos modelos hierárquicos com o objetivo de corrigir ambos os vieses. Contudo, pela falta de trabalhos teóricos acerca de um modelo desse tipo, optamos por ocultar esses resultados.

3.1 Estratégia Empírica

As próximas seções apresentam a estratégia empírica adotada para alcançar os nossos objetivos. Os modelos econométricos foram estimados utilizando as bases de dados discutidas nos subcapítulos anteriores, por meio do *software* estatístico STATA versão 11.2.

3.1.1 Estimação da Função de Produção Educacional

A regressão linear clássica faz uso do método de mínimos quadrados ordinários (MQO), que se baseia no princípio de que o estimador minimiza a soma do quadrado dos resíduos. O modelo populacional é linear nos parâmetros (Wooldridge, 2001, pág.49) e pode ser escrito, no nosso caso, como:

$$Y_i = X_i' \beta + u_i \quad (3.1)$$

$$u \sim N(0, \sigma^2) \quad (3.2)$$

A expressão acima descreve a nota e o vetor “X”, de dimensão $k \times 1$, como variáveis de cada candidato “i” cujos valores são observados em uma amostra aleatória da população e “ u_i ” é o erro aleatório não-observável. O vetor β , de dimensão $k \times 1$, denota os coeficientes de cada variável explicativa que devem ser estimados. As hipóteses básicas para que os coeficientes sejam identificados, e que minimizem o erro quadrático médio, podem ser escritas da seguinte maneira:

Hipótese 1: (Y_i, X_i) é independente e identicamente distribuída.

Hipótese 2: $E[X' u] = 0$

Hipótese 3: A matriz $E[X' X]$ possui posto completo

Com essas hipóteses, podemos estimar o vetor de estimadores $\hat{\beta}$, de tal forma que ele, admitindo-se as hipóteses 1, 2 e 3, é um estimador consistente para os

coeficientes populacionais. Se ainda, admitirmos a hipótese de exogeneidade estrita, então o estimador de MQO é não viesado:

$$\hat{\beta} = (X'X)^{-1} X'Y \quad (3.3)$$

3.1.2 A Correção do Viés de Auto-Seleção

Se a amostra observada não for aleatória, os coeficientes deixam de ser não-viesados. Para estes casos, existe um truncamento incidental da variável dependente, causado pela exclusão não aleatória de observações em nossa amostra (Greene, pág.780-781). O método mais utilizado na literatura, para corrigir a auto-seleção, é o de correção em dois estágios de Heckman (Heckman, 1979) e que será utilizado neste trabalho.

Os dados do ENEM apresentam esse truncamento, uma vez que a participação no exame é facultativa³. O método consiste em estimar duas regressões subsequentes, sendo a primeira, um modelo probit que calcula a probabilidade do indivíduo entrar na amostra. O segundo modelo consiste em estimar os coeficientes dos regressores que explicam a variável dependente por meio do método de mínimos quadrados ordinários, corrigido pela propensão do indivíduo não entrar na amostra, isto é, a razão inversa de Mills, calculada com base no primeiro estágio. A seguir o método será explicado em detalhes.

3.1.2.1 Metodologia da Correção do Viés de Auto-Seleção

Na primeira etapa da correção de Heckman, precisamos de informações tanto sobre os candidatos que não fizeram a prova, quanto daqueles que a prestaram. Nessa etapa, a variável dependente toma uma forma binária, com valor “1” se o indivíduo fez a prova e valor “0” se não fez a prova. Além disso, é necessário um

³ A participação dos estudantes no ENEM cresceu consecutivamente em todas as edições, o que pode atenuar o viés de auto-seleção.

vetor de variáveis explicativas para que possamos explicar a propensão de comparecer ao exame como função destas variáveis.

A abordagem da correção de Heckman pressupõe que as variáveis explicativas utilizadas no segundo estágio é um subgrupo daquelas utilizadas no primeiro estágio. Como a proposta desta dissertação é utilizar variáveis diferentes para cada período, ressaltamos que esta metodologia desvia-se, em parte, daquela originalmente proposta por Heckman (1979). No nosso caso, utilizamos apenas variáveis da escola, que não aparecem no segundo estágio.

Com as informações da escola, foi estimado um modelo probit para a probabilidade de o candidato prestar o exame:

$$p_i^* = Z_i' \gamma + u_{1i} \quad (3.4)$$

$$u_{1i} \sim N(0,1) \quad (3.4b)$$

Em que p_i^* é uma variável latente que representa a propensão do indivíduo em prestar o exame do ENEM, mas que não é observada. A informação observada é se o mesmo comparece ao exame ou não. A matriz Z_i representa as variáveis explicativas para a presença no exame, γ é o vetor de coeficientes.

Se $p_i^* > 0$, de modo que $u_{1i} > -Z_i' \gamma$, então $p_i = 1$. Dessa forma, considera-se que o indivíduo “i” compareceu ao exame.

Se $p_i^* \leq 0$, de modo que $u_{1i} \leq -Z_i' \gamma$, então $p_i = 0$. Dessa forma, considera-se que o indivíduo “i” **não** compareceu ao exame.

No caso da base de dados do ENEM, não temos informações sobre os jovens que não se inscreveram no exame, tampouco dos que se inscreveram e não compareceram à prova. Diferentemente, por exemplo, de informações sobre migração, mercado de trabalho ou fecundidade no Censo populacional, em que temos informação sobre todos os indivíduos.

Os dados não permitem fazer a correção de maneira direta, uma vez que necessitamos das informações dos jovens que não prestaram o exame para que possamos calcular a probabilidade da presença em função de variáveis observadas. Logo, a estratégia adotada para corrigir o viés de seleção consiste em três passos.

O primeiro passo foi calcular, para cada escola, a quantidade de alunos que estão no 3º ano do ensino médio e que compareceram ao exame, utilizando o questionário do ENEM 2008. Em seguida, calculou-se o número total de alunos no 3º ano do ensino médio em cada escola por meio do Censo Escolar 2008 e, por fim, esta informação foi inserida na base de dados do ENEM por meio da variável identificadora PK_COD_ENTIDADE, comum a ambas as bases, para então, calcular a proporção desejada.

No segundo passo, foram selecionadas aquelas escolas cuja participação calculada foi inferior a 10%, assumindo que todos os alunos dessa escola não fizeram a prova. Em seguida, selecionamos aquelas escolas cuja participação foi superior a 90%, assumindo presença no exame a todos os alunos dela⁴.

Com isso, atribuímos ausência na prova para algumas observações, ainda que, artificialmente, com valores válidos para as variáveis explicativas. Esse método pressupõe que as características de alunos provenientes de escolas com baixo percentual de presença sejam parecidas.

Com a amostra do primeiro estágio formada, estimou-se a probabilidade do candidato prestar o exame a partir dos dados da escola que frequenta, obtendo assim, a probabilidade estimada. Com isso, foi possível estimar a razão inversa de Mills, que mede o risco do indivíduo não ser selecionado. Heckman (1979, pág. 156) explica que a razão inversa de Mills é uma função monotonicamente decrescente da probabilidade de um indivíduo ser selecionado para a amostra.

⁴ A escolha da proporção de 10% e 90% foi baseada nos resultados de várias estimações do modelo de correção de Heckman com proporções diferentes. A comparação dos resultados mostrou que essa relação apresenta uma boa separação entre os grupos, sem que fossem perdidas muitas observações no segundo estágio.

O último passo foi estimar o segundo estágio, isto é, estimar a nota do ENEM em função de variáveis independentes e relevantes para a proficiência do estudante, representadas pela matriz X_i , considerando também, o erro do primeiro estágio. Para corrigir o viés utilizando o modelo do primeiro estágio, utilizou-se a razão inversa de Mills, denominada por λ_i , inserindo-a como uma variável explicativa no segundo estágio:

$$nota_i = X_i' \beta + \xi_i \quad (3.5)$$

$$\xi_i = \beta_m \lambda_i + u_{2i} \quad (3.6)$$

$$u_{2i} \sim N(0, \sigma^2) \quad (3.6b)$$

$$\lambda_i = \frac{\phi(Z_i' \hat{\gamma})}{\Phi(Z_i' \hat{\gamma})} \quad (3.7)$$

$$corr(u_{1i}, u_{2i}) = \rho \quad (3.8)$$

Onde $\phi(\cdot)$ e $\Phi(\cdot)$ representam, respectivamente, a função de densidade de probabilidade e função de probabilidade acumulada da distribuição normal. Com a introdução da razão inversa de Mills para corrigir o viés de seleção no segundo estágio, obtemos coeficientes não-viesados para os efeitos marginais da função de produção educacional.

Para calcular a correlação entre os erros do primeiro e segundo estágios, utilizamos o método proposto por Heckman (1979, pág.157), que estima o coeficiente de correlação consistente. Definindo SQR como a soma do quadrado dos resíduos do segundo estágio, temos que:

$$\delta_i = \lambda_i (\lambda_i - Z_i' \hat{\gamma}) \quad (3.9)$$

$$\hat{\sigma}^2 = \frac{SQR + \hat{\beta}_m^2 \sum_{i=1}^N \delta_i}{N} \quad (3.10)$$

$$\hat{\rho} = \frac{\hat{\beta}_m}{\hat{\sigma}} \quad (3.11)$$

3.1.3 A Violação da Hipótese de Amostras Independentes e Identicamente Distribuídas

A seguir, foram utilizados modelos hierárquicos lineares para estimar a função de produção educacional. A outra característica das bases de dados de educação que envia os coeficientes da regressão linear clássica é a sua estrutura hierárquica. As observações individuais, em sua maioria, não são totalmente independentes, como postula uma das hipóteses básicas do método de mínimos quadrados ordinários. A consequência disso é que a correlação média (ou intra-classe) entre variáveis, considerando estudantes de uma mesma escola, é maior que a correlação média entre variáveis considerando indivíduos de escolas diferentes (HOX, 2002, pág.05).

A violação da hipótese de independência dos dados faz com que as estimativas de erros-padrão dos testes estatísticos sejam menores, o que pode resultar em resultados significativos, porém espúrios.

O método dos modelos hierárquicos lineares considera o agrupamento não-aleatório dos dados, isto é, o fato de que os estudantes não estão igualmente distribuídos na população, pelo contrário, frequentam escolas muito distintas. Assim, não podemos ignorar o fato de que o desempenho de um aluno seja influenciado pelas características da escola que frequenta, e que tais fatores influenciam todos os alunos daquela escola. O uso de modelos de mínimos quadrados ordinários pode subestimar o efeito da escola sobre a proficiência (Albernaz et al., 2002).

A hipótese básica do modelo é que os dados possuem uma hierarquia, no sentido que os alunos de uma mesma escola fazem uso dos mesmos recursos

disponíveis, professores, etc. Para deixar mais claro ao leitor como essa hipótese é importante, considere que uma escola siga alguma linha pedagógica distinta das demais. Com isso, o aprendizado de todos os alunos dessa instituição é diferente das outras, podendo desenvolver atributos diferentes de outros jovens da mesma coorte.

3.1.3.1 Metodologia dos Modelos Hierárquicos Lineares

O modelo estimado foi composto por dois níveis, considerando as características do aluno no primeiro nível e as características da escola no segundo nível. O modelo multinível geral para o estudante “i” ($i=1, \dots, N_j$) em cada escola “j” ($j=1, \dots, J$), pode ser descrito como:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + u_{ij} \quad (3.12a)$$

Em que Y_{ij} é nota no ENEM do estudante “i” que frequenta a escola “j” e X_{ij} é a matriz de variáveis explicativas (nível 1) da nota do ENEM do estudante “i” que frequentou a escola “j”, e seus coeficientes são considerados fixos. Além disso, por hipótese:

$$u_{ij} \sim N(0, \sigma^2) \quad (3.12b)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + e_{0j} \quad (3.13a)$$

Onde:

$$\text{var}(e_{0j}) = \sigma_{00}^2 \quad (3.13b)$$

$$\text{cov}(e_{0j}, u_{ij}) = 0 \quad (3.13c)$$

Com o propósito investigativo, utilizou-se um procedimento de estimação em etapas, para que seja possível visualizar possíveis mudanças na variância explicada por cada nível, conforme o processo se desenrola. Em outras palavras,

primeiro estimou-se o modelo multinível sem variáveis explicativas para quaisquer níveis, com o objetivo de separar a variância devida à escola e ao aluno, e depois foram incluídas as variáveis de cada nível em modelos subsequentes⁵ (HOX, 1995). O modelo nulo (HLM0) ou ANOVA, sem variáveis explicativas, pode ser descrito como:

$$Y_{ij} = \gamma_{00} + e_{0j} + u_{ij} \quad (3.14)$$

O termo γ_{00} reflete a nota média da amostra inteira, e_{0j} significa o termo aleatório para cada escola “j” e o terceiro termo é o erro de cada indivíduo. A variância explicada pelo nível da escola é $\rho = \frac{\sigma_{00}^2}{\sigma_{00}^2 + \sigma^2}$ (variação intra-classe) e a variância explicada pelo nível do aluno (variação inter-classe) é igual a $1 - \rho = \frac{\sigma^2}{\sigma_{00}^2 + \sigma^2}$.

O próximo passo é estimar a função de produção educacional considerando as variáveis do aluno (modelo 1 ou HLM1) ou ANCOVA, assumindo coeficientes fixos:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + e_{0j} + u_{ij} \quad (3.15)$$

Após a inclusão das variáveis do candidato, podemos analisar a mudança no coeficiente de variação intra-classe, que poderá indicar se parte da variação que se observa entre escolas é, na verdade, reflexo das características pessoais e familiares de seus alunos. Para explicar parte da variação entre escola, foram incluídas variáveis explicativas da escola (modelo 2 ou HLM2):

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_j + \gamma_{10}X_{ij} + e_{0j} + u_{ij} \quad (3.16)$$

⁵ Poderíamos continuar o processo, atribuindo um termo aleatório ao coeficiente das variáveis, além do intercepto. Com isso o coeficiente seria uma função das variáveis de segundo nível e um efeito aleatório. No entanto, como o nosso objetivo não é investigar nenhuma variável específica, não prosseguimos por esta etapa.

Após a inclusão das variáveis de segundo nível, foi possível mais uma vez observar a diferença entre os coeficientes de correlação intra-classe e estudar como tais variáveis contribuem para explicar o desempenho dos alunos.

As estimações com dados empíricos serão apresentados no quinto capítulo e possuem um objetivo específico, o de analisar os determinantes da proficiência cognitiva. As seções a seguir buscam explicar a metodologia por trás do procedimento de análise do objetivo principal, a imputação da nota do ENEM na base de dados da PME e, conseqüentemente, a análise da proficiência imputada como determinante da situação ocupacional dos jovens brasileiros que estão concluindo o ensino médio.

3.1.4 Imputação da Nota Estimada do ENEM na PME

Para investigar a correlação entre a proficiência cognitiva, medida pelo desempenho na prova objetiva do ENEM, e a situação do indivíduo após a conclusão do ensino médio, utilizamos duas bases de dados distintas, o ENEM e a PME. A tentativa de unir duas bases diferentes é o maior dos desafios, uma vez que não podemos identificar o mesmo indivíduo nas duas bases de dados.

Para Soares (2010), é possível fazer um pareamento quase exato entre ENEM e PME utilizando a data de nascimento dos indivíduos, sexo, quantidade de pessoas na família e a região metropolitana. Sendo o ENEM quase censitário e a PME amostral, o autor forma a população-alvo com jovens da PME e busca indentificá-los na base do ENEM, mas essa tentativa reduz a amostra a ser explorada a cerca de menos de 1% da população que fez o ENEM entre 2003 a 2007. Na base da PME, restam em torno de 0,2% da base após o pareamento, pouco mais de 200 observações. Dessa forma, outra técnica que preserve um maior número de observações poderia ser explorada.

Outros autores utilizam o método de união de bases para estudar fenômenos educacionais e produtivos, como Curi e Menezes Filho (2006), em que a nota do exame de matemática do SAEB de 1995 é estimada em função de características pessoais e escolares disponíveis nesta base e na PNAD 1995 em um primeiro estágio. No segundo estágio, a nota estimada do SAEB, juntamente com o

resíduo da regressão da função educacional, foram imputadas na base do CENSO 2000, tomando os valores médios das variáveis por coorte.

Os autores fizeram uma correção do viés de auto-seleção utilizando a proporção de jovens nascidos entre 1977 e 1978 que tinham exatamente 10 anos de estudo em 1995 em relação ao total dessa coorte. No entanto, isso requer assumir certa homogeneidade da coorte, pois ocorre um agrupamento de pessoas que compartilham algum critério relativo a sua idade.

A metodologia de criação de mapas de pobreza desenvolvida por Elbers et al. (2003) utiliza uma pesquisa do tipo *survey* e um censo demográfico para associar informações de renda para regiões bastante desagregadas para, com isso, estimando índices de desigualdade regionais. Este tipo de metodologia leva em conta a riqueza de informações das pesquisas do tipo *survey* com a representatividade do censo. Ribas e Machado (2010) aplicaram este método para gerar estimativas de renda do não-trabalho e imputá-las na base de dados da PME, onde não consta esta informação.

O método desta etapa da dissertação segue aquela aplicada por Ribas e Machado (2010) e consiste, primeiramente, em estimar a nota da prova objetiva do ENEM⁶ como variável dependente de um vetor de covariáveis comuns a ambas as bases de dados. Logo, é essencial que sejam identificadas as variáveis disponíveis para explicar a proficiência em ambas as bases, utilizando a metodologia apresentada nas seções anteriores, em particular o método de mínimos quadrados ordinários.

Como esse método espera-se obter um resultado mais preciso do que se fosse utilizado o agrupamento por coortes, mas talvez não tão preciso quanto o pareamento quase-exato sugerido por Soares (2010), mas temos a vantagem de possuir mais observações do que ambas as alternativas.

⁶ A decisão de utilizar a nota da prova objetiva do ENEM, ao invés da nota da redação, foi baseada na comparação do coeficiente de determinação (R^2) dos modelos MQO estimados no quinto capítulo para a nota da prova objetiva e redação. O modelo para a prova objetiva mostrou-se superior, no que se refere ao ajuste dos dados, em relação ao modelo cuja variável dependente era a redação.

Os modelos estimados para a proficiência utilizando a correção de Heckman e os modelos Hierárquicos mostraram bastante semelhança entre si e com o MQO. Por isso, e também porque não temos como associar as variáveis escolares aos indivíduos da PME, utilizamos somente o modelo MQO:

$$profic = Q' \beta + e \quad (3.17)$$

$$e \sim N(0, \sigma^2) \quad (3.18)$$

A metodologia de Elbers et al. (2003) não assume que os erros são homoscedásticos e propõe que este parâmetro seja estimado conforme a equação a seguir:

$$\ln\left(\frac{e^2}{A - e^2}\right) = W' \alpha + \varepsilon \quad (3.19)$$

$$A = 1.05 \max(e^2) \quad (3.20)$$

Onde ε é homoscedástico, "A" é o limite superior da variância predita, α é o vetor de coeficientes e "W" é o vetor de variáveis explicativas. Com isso, o estimador do desvio-padrão, segundo os autores, é definido como:

$$\hat{\sigma} = \sqrt{\left(\frac{AB}{1+B}\right) + \frac{1}{2} \text{var}(\varepsilon) \left[\frac{AB(1-B)}{1-B^3}\right]} \quad (3.21)$$

$$B = \exp(W' \hat{\alpha}) \quad (3.22)$$

Dessa forma, o termo aleatório simulado é escrito como $\hat{e} \sim N(0, \hat{\sigma}^2)$ e, após a estimação deste parâmetro e do vetor de coeficientes $\hat{\beta}$, obtido pelo modelo de MQO, utilizamos as características observadas dos indivíduos " Q_i " em sua

primeira aparição na PME durante 2008 e 2009⁷, para estimar a sua proficiência cognitiva:

$$\widehat{\text{profic}}_i = Q_i' \hat{\beta} + \hat{e}_i \quad (3.23)$$

3.1.5 Estimação da Situação Ocupacional do Indivíduo em Função da Proficiência

O nosso último objetivo é utilizar os dados referentes à situação ocupacional dos jovens, na base de dados da PME, juntamente com a proficiência imputada, e investigar que papel ela desempenha no status ocupacional dos jovens após 12 meses.

A variável dependente aqui é a situação ocupacional do jovem após 12 meses da sua entrada na pesquisa mensal de emprego, sendo a proficiência cognitiva uma variável explicativa para ela. Ainda mais, foram utilizadas quatro amostras distintas, cada uma delas para a situação inicial declarada pelo indivíduo. Foram utilizados dados da primeira e quinta entrevista do indivíduo, de tal forma que o período entre elas é de doze meses.

No caso desta dissertação, foram construídas quatro situações para a variável dependente, estando os indivíduos sujeitos a alguma delas no período analisado: não trabalham e não estudam (**naotrabalha_ naoestuda**), só trabalham (**sotrabalha**), trabalham e estudam (**trabalha_estuda**) ou só estudam (**soestuda**). A metodologia utilizada para estimar a situação ocupacional baseia-se na estimação de um modelo logit multinomial, pois estamos interessados em analisar como o aumento da proficiência afeta a probabilidade de transitar de uma situação para outra.

Os determinantes da transição ocupacional também já foram investigados no Brasil por Curi e Menezes-Filho (2004) por meio do modelo logit multinomial,

⁷ Onde " Q_i " é um subvetor de " X_i ", correspondente às variáveis utilizadas no quinto capítulo para explicar a proficiência dos estudantes, conforme características pessoais e familiares que também estão disponíveis na PME.

utilizando a PME de 1984 a 2001, para investigar a mobilidade dos trabalhadores entre as categorias de trabalho formal, informal, conta própria, desempregado e inativo. Em outro trabalho, Flori e Menezes-Filho (2008), estimam um modelo logit multinomial para estimar a probabilidade de transição entre ocupações, controlando por um vetor de variáveis pessoais, utilizando a PME de 1984-1985 e 2000-2001.

Os modelos estimados serão divididos seguindo a metodologia de Curi e Menezes-Filho (2004) e Flori e Menezes-Filho (2008), que estimam um modelo logit multinomial para cada subamostra composta pelas pessoas com a mesma situação inicial.

Como estamos considerando apenas os jovens que estão no último ano do ensino médio, existem, inicialmente, apenas duas situações possíveis, trabalhar e estudar, ou apenas estudar. Por isso, são possíveis dois modelos multinomiais, mas como também foram estimados modelos separados para homens e mulheres, totalizando quatro modelos estimados.

Em nossa análise, utilizamos dados de indivíduos que estão cursando o 3º ano do ensino médio. Além disso, foram utilizadas as bases de dados de 2008 a 2010, incluindo todos os meses, com a finalidade de garantir mais observações para as nossas estimações. Assim, observamos a situação inicial dos jovens que entraram na amostra em 2008, e doze meses depois, em 2009. Da mesma forma, acompanhamos os jovens que entraram na amostra em 2009 e depois de doze meses, em 2010.

A probabilidade de um dado indivíduo “i” pertencer a uma categoria “j”, com X_i denotando o vetor de variáveis explicativas do indivíduo, pode ser escrita como:

$$p(y = j | x) = \frac{e^{x_i \beta_j}}{1 + \sum_{h=1}^3 e^{x_i \beta_h}} \quad (3.24)$$

A soma das probabilidades deve ser igual a unidade e, portanto, assumimos que uma das categorias possui o coeficiente igual a zero, que será a chamada de

categoria de referência “r”. Segundo Wooldridge (2001 p. 498), modelo é estimado por máxima verossimilhança e, para cada indivíduo “i”, a função de log-verossimilhança é dada pela equação 3.25. Os coeficientes são identificados ao maximizar a somatória da função abaixo, considerando todos os indivíduos.:

$$l_i = \sum_{j=0}^J 1[y_i = j] \log[p_j(x_i, \beta)] \quad (3.25)$$

O principal objetivo é estudar o efeito parcial da proficiência sobre a probabilidade de transição, segundo cada categoria de referência. No entanto, os resultados são discutidos com referência à razão de risco relativo.

A razão de risco relativo (RRR) fornece uma informação sintética sobre o efeito marginal da variável de interesse de forma relativa. Com a razão de risco relativo, podemos verificar se, para o aumento de um ponto na escala de proficiência cognitiva do indivíduo, ele estará mais propenso a mudar para uma determinada categoria ou a permanecer na categoria de referência. A princípio notamos que a probabilidade relativa, ou risco relativo, de uma categoria, em relação à categoria de referência é dada por:

$$\frac{\Pr(y = j | x)}{\Pr(y = r | x)} = e^{x\beta_j} \quad (3.26)$$

O interesse principal está em saber se um aumento da proficiência cognitiva aumenta as chances de um indivíduos mudar de posição no período em questão. Assim, para um aumento de um ponto na escala de proficiência, temos que a razão de risco relativo é dada por:

$$RRR = \frac{\Pr(y = j | x + 1)}{\Pr(y = r | x + 1)} \bigg/ \frac{\Pr(y = j | x)}{\Pr(y = r | x)} = e^{\beta_j} \quad (3.27)$$

A razão de risco relativo compara o risco relativo de transição, para um dado aumento na proficiência. Assim, se ela for maior que a unidade, então o risco relativo do indivíduo mudar da categoria de referência “r” para a categoria “j”, no período de 12 meses, aumenta com a sua proficiência cognitiva. Por outro lado,

se a razão for menor que a unidade, então o risco relativo de mudar de categoria diminui conforme a proficiência cognitiva aumenta.

Os resultados da estimação do modelo multinomial são apresentados no capítulo 4, bem como a análise da razão de risco relativo da proficiência cognitiva como determinante da probabilidade do indivíduo transitar entre situações ocupacionais. Com isso, o objetivo da nossa análise é concluído e a discussão teórica, levantada no primeiro capítulo, juntamente com a análise dos determinantes da proficiência cognitiva, apresentados no segundo capítulo, propiciam mais uma visão sobre o processo de acumulação de capital humano.

3.2 Apresentação das Bases de Dados e Variáveis Utilizadas na Estratégia Empírica

A função de produção educacional foi estimada utilizando a base de dados do ENEM de 2008 e o primeiro estágio da correção de Heckman utilizou, além do ENEM, o Censo Escolar de 2008, para estimar a probabilidade do candidato comparecer ao exame. Para investigar a situação ocupacional dos jovens após a conclusão do ensino médio, utilizamos a Pesquisa Mensa de Emprego (PME) de 2008 a 2010.

Esta seção apresenta as principais características dessas bases de dados para, posteriormente, discutir as estatísticas descritivas das variáveis utilizadas nos capítulos posteriores.

3.2.1 Descrição da Base de Dados do ENEM e das Variáveis da Função de Produção Educacional

No Brasil, existem muitos trabalhos sobre capital humano e uma vasta quantidade de bases de dados sobre atributos socioeconômicos e demográficos de seus habitantes produzidos pelo IBGE (Instituto Brasileiro de Geografia e Estatística) e bases de dados referentes ao sistema educacional fornecidos pelo INEP (Instituto Nacional de Pesquisas Educacionais Anísio Teixeira).

Os trabalhos sobre desempenho escolar são, em geral, uma aplicação da função de produção educacional no Brasil com base no exame do SAEB, disponível para a 4ª e 8ª série do ensino fundamental e 3º ano do ensino médio (Barbosa e Fernandes, 2001; César e Soares, 2001; Albernaz et al., 2002; Soares, 2004; Franco et al., 2007) ou outros exames regionais, como o SIMAVE em Minas Gerais (Machado et al., 2005; Soares, 2005).

O exame do ENEM, por sua vez, fornece dados importantes para investigação dos determinantes do desempenho educacional dos jovens brasileiros. Segundo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), o ENEM pode ser descrito como:

[...] um exame individual e de caráter voluntário, oferecido anualmente aos concluintes e egressos do ensino médio, com o objetivo principal de possibilitar uma referência para auto-avaliação, a partir das competências e habilidades que o estruturam. Além disso, ele serve como modalidade alternativa ou complementar aos processos de seleção para o acesso ao ensino superior e ao mercado de trabalho. (BRASIL, 2005)

Esse exame também diferencia-se do Saeb por sua proposta de mensurar competências, ao invés de habilidades propriamente ditas. A diferença entre estes conceitos, por exemplo no caso da leitura, saber ler é uma habilidade, mas interpretar um texto de forma lógica é considerada uma competência. Assim, de acordo com BRASIL (2008), são cinco as competências cerceadas pelo ENEM:

1. Competência I: Dominar a norma culta da língua portuguesa e fazer uso das linguagens matemática, artística e científica;
2. Competência II: Construir e aplicar as dimensões de várias áreas do conhecimento para a compreensão de fenômenos naturais, de processos histórico-geográficos, de produção tecnológica e das manifestações artísticas;
3. Competência III: Selecionar, organizar, relacionar e interpretar dados e informações representados de diferentes formas, para tomar decisões e enfrentar decisões-problema;

4. Competência IV: Relacionar informações, representadas de diferentes formas, e conhecimentos disponíveis em situações concretas, para construir argumentação consistente;
5. Competência V: Recorrer aos conhecimentos desenvolvidos na escola para elaboração de propostas de intervenção solidária na realidade, respeitando os valores humanos e considerando a diversidade sócio-cultural.

O ENEM fornece um retrato da proficiência cognitiva de muitos estudantes que possuem o objetivo de ingressar no ensino superior e também daqueles que anseiam por entrar no mercado de trabalho. Assim, fornece informações ao fim do ensino básico para os brasileiros.

A nota da redação do ENEM depende da proficiência que os candidatos possuem sobre as competências de dominar a norma culta da língua escrita, aplicar conceitos de várias áreas de conhecimento dentro da estrutura do texto, organizar e interpretar fatos para defesa de um ponto de vista, construir argumentos e elaborar uma proposta de intervenção ao problema apresentado.

A análise da proficiência também será feita com base na nota de redação do ENEM. Essa abordagem visa, principalmente, confirmar o resultado encontrado por Golgher (2010a), que encontrou uma vantagem em proficiência para o sexo feminino na prova de redação. Ademais, como esse teste possui o mesmo intuito que a prova objetiva, espera-se obter efeitos similares para a determinação da proficiência, apesar de que a liberdade de expressão na redação pode resultar alguma diferença.

Os dados mais recentes, na data em que a análise foi feita, eram referentes ao ENEM de 2008, e como o exame é voluntário, apesar de termos 4.018.070 inscrições, muitas não podem ser naturalmente utilizadas em nossa análise, pois os candidatos não fizeram a prova e/ou não responderam o questionário socioeconômico. Dessa forma, e como estamos interessados em estudar os jovens que estão terminando o ciclo básico da educação, foram excluídos todos que não responderam o questionário socioeconômico e que não tinham previsão de terminar o ensino médio em 2008.

Para concentrar a investigação apenas nos jovens, foram excluídos aqueles que estão muito acima da média de idade para cursar o último ano do ensino médio, bem como aqueles muito jovens, restringindo a amostra àqueles cuja idade está entre 15 e 24 anos na época do exame, o que caracteriza mais de 80% daqueles candidatos com previsão de terminar o ensino médio nesse ano.

A diferença de proficiência entre diferentes sexos será captada por uma variável indicadora se o candidato é do sexo feminino (**mulher**). A cor/raça do indivíduo também pode influenciar o seu desempenho, ainda mais quando há uma concentração de algumas delas em diferentes camadas sociais, e por esse motivo, a variável (**nao_branco**) indica se o candidato declarou-se preto/pardo, ao invés de branco/amarelo.

Além dessas variáveis pessoais, utilizamos outras que podem ser importantes determinantes da proficiência, em razão da alocação do tempo do estudante. A primeira delas é se ele estuda no período noturno (**noturno**) e a segunda é se ele trabalha (**trabalha**).

Assim como em Golgher (2010b, pág.21) e Andrade e Laros (2007), a variável indicadora se o candidato trabalha foi incluída na análise por explicar parte do prejuízo sobre o aprendizado do estudante, principalmente por causa do menor tempo disponível aos estudos e também, pode indicar a carência de recursos familiares que possibilitem ao aluno dedicar-se inteiramente aos estudos.

Com respeito aos determinantes escolares, utilizamos dummies para a rede administrativa da escola em que o estudante cursa o ensino médio. A variável de referência é a da rede estadual (**esc_estad**) e as outras são: rede municipal (**esc_munic**), rede federal (**esc_fed**) e rede privada (**esc_priv**). Além disso, controlamos pela variável *dummy* indicativa do estudante frequentar curso preparatório de vestibular (**cursinho**).

Ao considerar a estrutura familiar e seus recursos destinados à educação de seus filhos, consideramos identificar o efeito de dividir o domicílio com irmãos (**irmaos_m**), por justificar a alocação e diluição de tais recursos entre os filhos. Esta variável pode apresentar um viés da forma como ela é construída, pois o questionário do ENEM pergunta se o candidato mora com irmãos, mas não se ele

possui irmãos, os quais poderiam não compartilhar o mesmo domicílio, mas dividir os recursos mesmo assim. Por fim, podemos considerar também o número de moradores dentro do domicílio, porque os recursos destinados à educação podem diluir-se à medida que ela não é a única demanda familiar.

No âmbito familiar, as duas outras variáveis utilizadas foram a renda familiar e a escolaridade da mãe. A renda familiar inclui a renda obtida de todas as fontes dos moradores do domicílio, e esta dividida em faixas: o domicílio não possui renda (**renda0sm**), renda até 1 salário mínimo (**renda1sm**), renda de 1 a 2 salários mínimos inclusive (**renda1a2sm**), renda de 2 a 5 salários mínimos inclusive (**renda2a5sm**), renda de 5 a 10 salários mínimos inclusive (**renda5a10sm**) e renda acima de 10 salários mínimos (**renda10sm**).

A escolaridade da mãe foi dividida também em ciclos, pois o questionário do ENEM não pergunta a escolaridade em anos de ensino. Logo, temos diversas *dummies* que informam o nível máximo de escolaridade da mãe: se não frequentou a escola (**mae_naoestudou**), se desconhece a escolaridade da mãe (**mae_naosabe**), 1ª a 4ª série do ensino fundamental (**mae_fund1a4**), 5ª a 8ª série do ensino fundamental (**mae_fund5a8**), ensino médio incompleto (**mae_em1**), ensino médio completo (**mae_em2**), ensino superior incompleto (**mae_sup1**), ensino superior completo (**mae_sup2**) e pós-graduação (**mae_pos**).

O sucesso escolar do aluno é um indicador central do seu desempenho escolar, e assim, incluiu-se uma variável da idade do candidato (**idade**) centralizada na idade considerada normal para quem está no 3º ano do ensino médio (17 anos). Essa variável se justifica por sintetizar o baixo aproveitamento que o estudante em atraso escolar e que sinaliza dois problemas recorrentes no sistema educacional, as taxas de abandono e repetência.

A restrição da análise apenas às regiões metropolitanas também implica a desconsideração daqueles que se declararam indígenas, por representarem uma proporção muito pequena da amostra e não pertencerem ao foco deste trabalho, que é a compatibilização das bases de dados do ENEM e PME. Após todas as exclusões, consideramos apenas candidatos residentes das seis regiões metropolitanas presentes na PME (Salvador, Recife, Porto Alegre, Belo Horizonte,

São Paulo, Rio de Janeiro), representadas, respectivamente, pelas variáveis: **reg_sa**, **reg_re**, **reg_pa**, **reg_bh**, **reg_sp** e **reg_rj**. Com isso, ficamos com 159.104 observações válidas. Elas incluem os candidatos brancos, negros, amarelos e pardos, que vivem em uma das seis regiões metropolitanas acima mencionadas, que estão na faixa etária de 17 a 24 anos e que responderam ao questionário socioeconômico.

3.2.2 Descrição da Base de Dados do Censo Escolar e das Variáveis Utilizadas na Correção de Heckman

O Censo Escolar é um levantamento feito anualmente pelo Instituto Nacional de Estudos e pesquisas Educacionais Anísio Teixeira (INEP) e abrange todas as instituições de educação básica de todas as redes administrativas. As informações que constam na pesquisa se referem, em linhas gerais, a infraestrutura escolar, tipo de ensino oferecido, características pessoais dos alunos, características pessoais dos docentes e a sua escolaridade, e também, informações referentes a turmas. Como já foi explicado na seção **3.1.2**, utilizamos as variáveis do censo escolar no durante o processo de correção do viés de auto-seleção.

O primeiro estágio da correção de Heckman estima a probabilidade do indivíduo entrar na amostra e, para isso, selecionamos aquelas variáveis que julgamos importantes para determinar a qualidade das escolas. Como estamos tratando de escolas situadas em regiões metropolitanas, as escolas possuem um nível de urbanização bastante elevado, de tal maneira que todas possuem uma infraestrutura de saneamento básico adequado. Por isso, não convém explorar características básicas de funcionamento, como eletricidade, esgoto, água, etc.

Para o nosso objetivo, selecionamos variáveis referentes a docentes, importantes na diferenciação da qualidade das escolas, em particular, a idade média dos docentes de cada escola e a proporção de docentes na escola com alguma pós-graduação.

Em relação às variáveis de infraestrutura educacional, identificamos as seguintes: presença de laboratório de informática, presença de laboratório de ciências,

biblioteca, acesso à internet para os alunos e presença banda larga. A análise descritiva detalhada dessas variáveis será feita no próximo capítulo.

3.2.3 Descrição da Base de Dados da PME Utilizada para a Estimação da Situação Ocupacional dos Indivíduos

A seguir são apresentadas as variáveis utilizadas para a imputação da nota do ENEM na base de dados da PME. Mas antes, devemos fazer algumas considerações sobre a Pesquisa Mensal de Emprego.

A PME possui o objetivo de produzir indicadores sobre o mercado de trabalho de seis regiões metropolitanas: Salvador, Recife, Porto Alegre, Belo Horizonte, Rio de Janeiro e São Paulo. As informações são coletadas de cada indivíduo, sendo este a unidade de investigação, que mora em uma unidade domiciliar selecionada (BRASIL, 2007). Logo, caso uma mesma família permaneça no domicílio por tempo suficiente, podemos observar o mesmo indivíduo em dois momentos distintos.

Alguns problemas de atrito são encontrados nesta base de dados e resumem-se a três principais: a migração de domicílio, a recusa de entrevista e a resposta inadequada em diferentes entrevistas posteriores (o que impossibilita identificar com exatidão o mesmo indivíduo em diversos períodos de tempo).

Para diminuir a perda de informação por atrito, foi utilizado um algoritmo de emparelhamento avançado, proposto por Ribas & Soares (2008), cujo objetivo é reduzir a taxa de “falso atrito”, devido à imprecisão das informações fornecidas por um domicílio, que é uma das causas de perda de observações na base de dados. Assim, o método sugerido pelos autores, que consiste em buscar a mesma pessoa em diferentes períodos de tempo, foi utilizado com a finalidade de melhorar a qualidade das informações discutidas neste trabalho.

Uma outra característica da base de dados da PME é o seu desenho amostral. Isso significa que a forma como ela é montada pode atribuir diferentes pesos para cada observação, uma vez que cada domicílio não possui a mesma probabilidade de ser selecionado:

A PME utiliza uma amostra probabilística de domicílios, estratificada e conglomerada em dois estágios, para cada região metropolitana de abrangência da pesquisa. Na PME, os municípios e pseudomunicípios (conjuntos de municípios de menor porte em quantidade de domicílios segundo o Censo Demográfico 2000) correspondem, cada um, a estratos independentes de seleção, garantindo assim o espalhamento da amostra pela região metropolitana.

Dentro de cada município ou pseudomunicípio é feita a seleção das unidades primárias de amostragem - UPAs e, posteriormente, das unidades secundárias de amostragem - USAs. As unidades primárias de amostragem da pesquisa são os setores censitários, enquanto as unidades secundárias de amostragem são as unidades domiciliares. (BRASIL, 2007, pág. 28-29)

A estratificação da amostra induz ao uso de pesos para estimar parâmetros populacionais, porque a distribuição das variáveis pode diferir entre os subgrupos, de forma que a hipótese das amostras identicamente distribuídas falha. No entanto, a estratificação da PME não foi feita com base na variável dependente que utilizamos. Por isso e com base em Ciol *et al.* (2006), e pelo fato de não estirmarmos um parâmetro regional que represente a população, não utilizamos a expansão das observações⁸.

As variáveis disponíveis na PME são, em sua maioria, referentes à posição e rendimento no mercado de trabalho, e poucas se referem a características pessoais dos indivíduos. Apesar disso, as variáveis pessoais disponíveis permitem estimar a proficiência cognitiva, que é o nosso objetivo.

A PME é composta por diferentes grupos rotativos, dentro dos quais são observados os domicílios de áreas censitárias de seis regiões metropolitanas brasileiras, a saber: São Paulo, Rio de Janeiro, Porto Alegre, Belo Horizonte, Recife e Salvador. A coleta das informações é feita, então, com os moradores destes domicílios.

Para que a nossa análise seja feita de maneira parcimoniosa, optamos por utilizar apenas alguns desses grupos rotativos. Ao considerar apenas os estudantes que

⁸ Para fins de investigação, o modelo logit multinomial foi estimado com a amostra expandida, e comprovamos que os coeficientes não apresentaram nenhuma alteração em seu sentido. No entanto, como a expansão multiplica os casos existentes, muitos coeficientes tornaram-se estatisticamente significativos.

estão concluindo o ensino médio na base de dados do ENEM 2008, a preocupação inicial foi selecionar na PME os indivíduos tal que as amostras de ambas as bases fosse o mais homogêneo possível.

A imputação da nota do Enem foi feita com base na prova realizada em agosto de 2008, considerando apenas os estudantes que estão concluindo o ensino médio nesse ano. Sendo assim, buscaram-se indivíduos cuja primeira entrevista na PME aconteceu entre janeiro de 2008 e dezembro de 2009, isto é, consideramos que a proficiência cognitiva estimada por meio dos indivíduos que prestaram o exame em agosto de 2008, na base do ENEM, pode ser atribuída àqueles que estão cursando o último ano do ensino médio em 2008 e 2009, na base da PME, sem perda de generalidade.

A hipótese é que a proficiência cognitiva, da maneira como ela foi estimada, não possui grandes variações no curto prazo, tornando possível imputar um nível de proficiência para os estudantes de 2009 com base nas informações de 2008.

Com o intuito de adicionar mais observações e diminuir o atrito que ocorre ao longo de um ano na PME, devido à mudança de domicílio por parte das pessoas que respondem à pesquisa, optou-se, também, por utilizar os grupos rotacionais que entraram na pesquisa janeiro de 2008 a dezembro de 2009, respectivamente, os grupos rotativos L4 até O3.

Ademais, como estamos interessados em estudar a influência da proficiência cognitiva dos jovens sobre a inserção no mercado de trabalho no prazo de 12 meses, foram excluídos os indivíduos com menos de 17 anos ou maiores de 24 anos, que representavam 9% das 2019 observações, restando 1838 casos. O quadro 2 mostra o número de observações válidas, de acordo com o mês e ano em que o indivíduo entrou na amostra da PME⁹:

⁹ Foi utilizado apenas a primeira e quinta entrevista de cada indivíduo, de forma que observamos sua situação quando entrou na amostra e doze meses após a entrada.

QUADRO 2 - Número de indivíduos por mês e ano de entrada na PME

Mês	Ano		Total
	2008	2009	
Janeiro	60	72	132
Fevereiro	89	69	158
Março	70	74	144
Abril	92	84	176
Maio	92	60	152
Junho	82	82	164
Julho	74	77	151
Agosto	87	83	170
Setembro	100	82	182
Outubro	78	80	158
Novembro	69	76	145
Dezembro	61	45	106
Total	954	884	1.838

Fonte dos dados básicos: IBGE - PME 2008 e 2009

Em primeiro lugar, investigamos esse grupo de indivíduos e como eles estão inseridos no mercado de trabalho (trabalhando ou procurando emprego) e se estão estudando. Descartamos de nossa análise aqueles que repetiram o último ano do ensino médio e consideramos, na mesma categoria, os jovens que estão no primeiro ano da faculdade ou frequentando um curso pré-vestibular, porque, apesar de não estarem no ensino superior após o término do ensino médio, mostram-se interessados em continuar sua educação formal.

A pessoas foram distribuídas em duas situações iniciais. A primeira situação é aquela em que o jovem está ativo no mercado de trabalho e está estudando, a outra ocorre quando o jovem está apenas estudando. As situações finais são divididas em quatro diferentes situações: não estuda e não está ativo no mercado de trabalho (1), está ativo no mercado de trabalho e não estuda (2), está ativo no mercado de trabalho e estuda (3), estuda e não está ativo no mercado de trabalho (4).

4 ANÁLISE DESCRITIVA DAS VARIÁVEIS UTILIZADAS NA ESTRATÉGIA EMPÍRICA

Neste capítulo, apresentamos as estatísticas descritivas das variáveis do ENEM 2008 utilizadas para estimar a função de produção educacional, juntamente com aquelas variáveis do Censo Escolar 2008 utilizadas no primeiro estágio da correção de Heckman.

Além disso, comparamos as variáveis comuns às bases de dados do ENEM 2008 e PME 2008-2009, que foram utilizadas na etapa de imputação da proficiência. Por fim, descrevemos a variável de situação ocupacional disponível na PME que será usada para investigar a importância da proficiência cognitiva como determinante da mobilidade ocupacional dos jovens.

4.1 Análise Descritiva das Variáveis Utilizadas na Estimação da Função de Produção Educacional

As variáveis utilizadas nessa etapa foram apresentadas no terceiro capítulo e a Tabela 1 mostra um resumo das principais estatísticas descritivas das variáveis que foram usadas para estimar a função de produção educacional, para as 159.104 observações utilizadas inicialmente na aplicação do MQO. Sobre a descrição destes resultados, em primeiro lugar, a variável que identifica o sexo demonstra que 61,1% dos indivíduos é do sexo feminino e a variável de cor/raça mostra que 44,7% declarou ser negro/pardo. A distribuição dessas duas variáveis dentro de cada região metropolitana será discutida adiante.

Os indivíduos que trabalham, enquanto estudam no 3º ano do ensino médio, representam 23,8% da amostra. A proporção de candidatos que trabalham é baixa entre aqueles cuja renda familiar é mais alta. Isto é, a condição de trabalho entre os estudantes é mais presente nas famílias menos favorecidas. As regiões metropolitanas que concentram uma maior proporção de jovens que trabalhavam enquanto estudavam são as do Sul e Sudeste do país, respectivamente, Porto Alegre (31,9%), São Paulo (30%) e Belo Horizonte (24,8%).

TABELA 1 – Resumo das estatísticas descritivas de todas as variáveis utilizadas na estimação da função de produção educacional

Variável	Média	Desvio-Padrão	Mínimo	Máximo
nota	46,72	16,51	0	98,41
idade	0,768	1,17	-2	7
mulher	0,611	0,49	0	1
nao_branco	0,447	0,50	0	1
trabalha	0,238	0,43	0	1
irmaos_m	0,758	0,43	0	1
noturno	0,241	0,43	0	1
cursinho	0,172	0,38	0	1
esc_munic	0,023	0,15	0	1
esc_fed	0,017	0,13	0	1
esc_priv	0,288	0,45	0	1
renda1sm	0,093	0,29	0	1
renda1a2sm	0,289	0,45	0	1
renda2a5sm	0,345	0,48	0	1
renda5a10sm	0,140	0,35	0	1
renda10sm	0,125	0,33	0	1
mae_fund1a4	0,158	0,37	0	1
mae_fund5a8	0,183	0,39	0	1
mae_em1	0,085	0,28	0	1
mae_em2	0,280	0,45	0	1
mae_sup1	0,055	0,23	0	1
mae_sup2	0,147	0,35	0	1
mae_pos	0,055	0,23	0	1
moradoresXrenda1sm	0,360	1,21	0	7
moradoresXrenda1a2sm	1,188	2,00	0	7
moradoresXrenda2a5sm	1,451	2,12	0	7
moradoresXrenda5a10sm	0,573	1,48	0	7
moradoresXrenda10sm	0,507	1,39	0	7
reg_sa	0,062	0,24	0	1
reg_re	0,079	0,27	0	1
reg_rj	0,198	0,40	0	1
reg_sp	0,470	0,50	0	1
reg_pa	0,063	0,24	0	1

Fonte: Elaboração própria com dados do ENEM 2008

A variável indicativa “cursinho” mostra que 17% frequenta algum tipo de curso preparatório para o vestibular e/ou ENEM. Mas como esses cursos geralmente são privados, as famílias mais pobres podem não conseguir arcar com o seu custo. Por isso, observamos que a frequência de candidatos que frequentam um curso preparatório pertencem também a famílias com renda mais elevada.

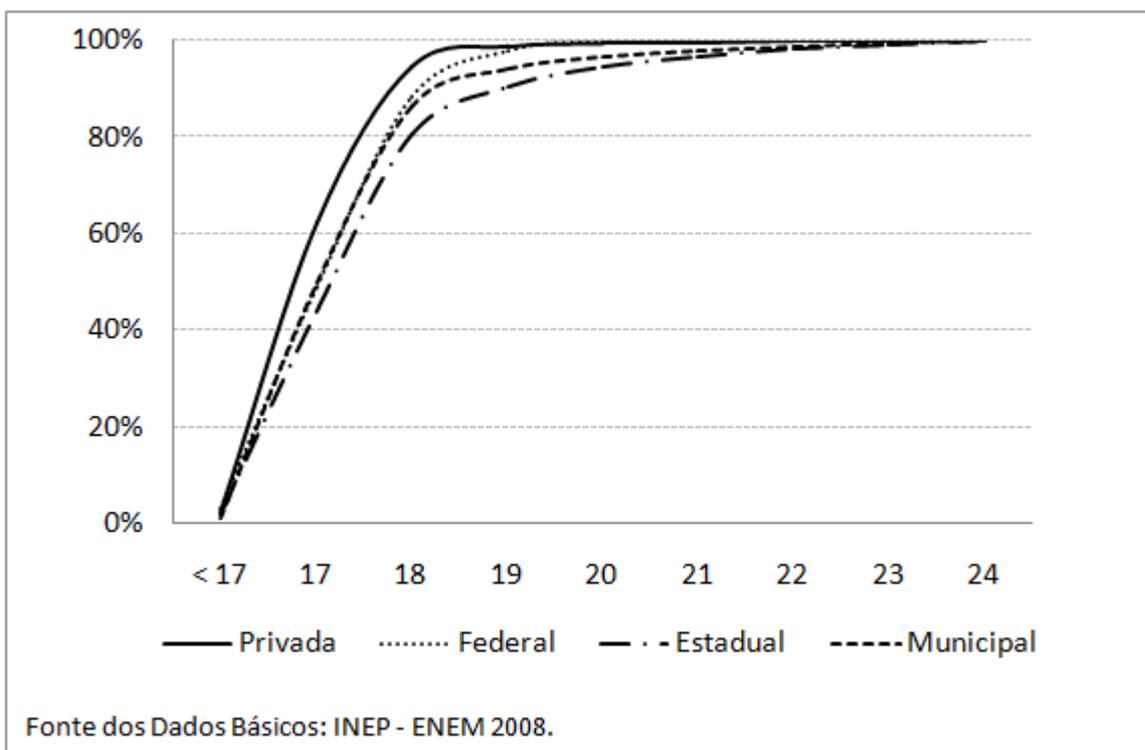
A variável indicativa “residir com irmãos” mostra que 75,8% dos jovens mora com irmãos, e ela foi usada para estimar o efeito de dividir recursos econômicos entre

eles. Na análise da variável por faixa de renda familiar, notamos não haver grande diferença entre famílias ricas e pobres.

Os estudantes do período noturno somam 24% da amostra, e a proporção destes que também trabalham é de pouco mais da metade (52%). Assim, uma grande parcela dos estudantes do optaram pelo período noturno, mas estão ocupando o dia com outras atividades que não o trabalho. Além disso, a relação entre estudar à noite e a renda familiar é inversamente proporcional, com pouquíssimos casos para estudantes na faixa de renda mais elevada.

A idade do estudante é mais um fator preponderante na determinação da sua proficiência cognitiva, cabendo assim, analisar como é a sua distribuição na amostra aliada a outras variáveis. Em primeiro lugar, identificamos que a rede privada concentra menos estudantes com atraso escolar, em comparação às redes públicas. O Gráfico Nº 1 mostra como a diferença de idade se distribui de forma diferenciada conforme a rede administrativa da escola que o aluno frequenta:

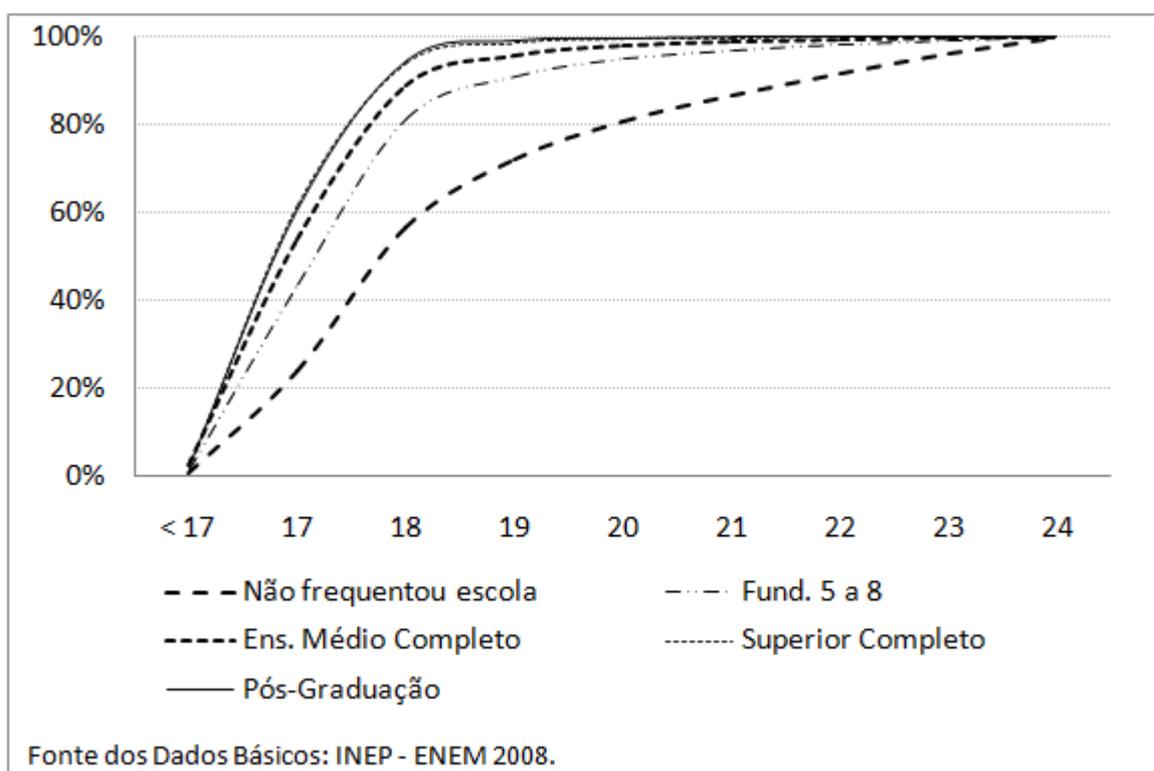
GRÁFICO 1 - Distribuição acumulada da idade dos candidatos por rede de ensino



A idade avançada no final do ensino médio também pode apontar possíveis restrições que o jovem enfrenta durante sua vida escolar. Como já foi explorado anteriormente, a família proporciona ao jovem capital econômico e cultural para sua formação educacional, e a presença destes atributos em quantidades pequenas pode ter um impacto negativo sobre o seu aproveitamento.

Dessa forma, considerando que a escolaridade da mãe é um fator positivamente correlacionado com o nível socioeconômico familiar, podemos ver pelo Gráfico N° 2 que o atraso escolar é maior entre aqueles cuja família possui condições sociais mais precárias, notando que a relação entre escolaridade da mãe e atraso escolar possui uma relação não-linear.

GRÁFICO 2 - Distribuição acumulada da idade dos candidatos que estão no 3º ano do ensino médio segundo a escolaridade da mãe.



A seguir descrevemos a variável de renda familiar, que segue a resposta fornecida pelo questionário e indica a renda familiar do candidato por faixas de salários mínimos, como pode ser observada no Tabela N° 2: se a família não possui renda (**renda0sm**), renda familiar até 01 salário mínimo inclusive

(renda1sm), renda familiar de 01 a 02 salários mínimos inclusive **(renda1a2sm)**, renda familiar de 02 a 05 salários mínimos inclusive **(renda2a5sm)**, renda familiar de 05 a 10 salários mínimos inclusive **(renda5a10sm)** ou renda familiar acima de 10 salários mínimos **(renda10sm)**.

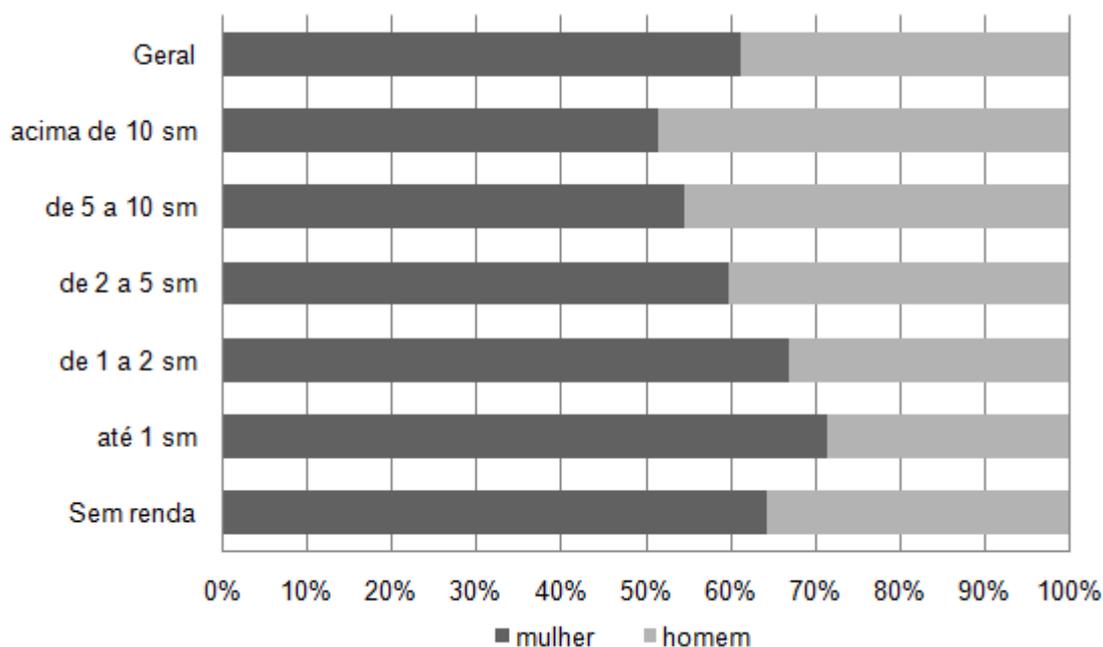
TABELA 2 – Distribuição total e por região metropolitana dos candidatos segundo a renda familiar

	Salvador	Recife	Porto Alegre	Belo Horizonte	Rio de Janeiro	São Paulo	Total
Sem Renda	1,2%	1,7%	0,5%	0,4%	0,7%	0,6%	0,7%
Até 1 salário mínimo	15,6%	23,9%	7,1%	7,0%	11,4%	5,7%	9,1%
De 1 a 2 s.m.	33,6%	29,5%	32,9%	30,4%	30,6%	26,2%	28,7%
De 2 a 5 s.m.	26,7%	21,5%	40,8%	39,7%	30,3%	37,6%	34,7%
De 5 a 10 s.m.	10,5%	11,3%	12,2%	12,7%	13,8%	15,8%	14,1%
Acima de 10 s.m.	12,5%	12,0%	6,5%	9,9%	13,2%	14,2%	12,7%
Total	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
Observações	9.688	12.573	10.147	20.184	31.538	74.974	159.104

Fonte dos dados básicos: ENEM 2008

O gráfico abaixo mostra uma diferença nítida no que diz respeito às alternativas educacionais das mulheres e homens segundo o seu nível sócioeconômico. Podemos observar a maior proporção de mulheres prestando o ENEM em famílias cuja renda domiciliar é pequena, fato que poderia ser influenciado pela decisão dos homens de famílias mais pobres de antecipar a entrada no mercado de trabalho.

GRÁFICO 3 - Proporção de candidatos ao ENEM segundo gênero e faixa de renda familiar



Fonte dos Dados Básicos: INEP - ENEM 2008

Apesar de muitos trabalhos utilizarem anos de estudo como variável, a base de dados do ENEM não permite essa construção devido à pergunta feita ao candidato. Posto isso, e dado que a escolaridade do pai e da mãe são positivamente correlacionadas, optou-se por utilizar somente a educação da mãe da forma como foi apresentada no capítulo anterior. A distribuição dos estudantes segundo a escolaridade da mãe pode ser vista na Tabela Nº 3:

TABELA 3 – Distribuição dos candidatos por RM segundo escolaridade da mãe

	Salvador	Recife	Porto Alegre	Belo Horizonte	Rio de Janeiro	São Paulo	Total
Não Estudou	2,1%	2,2%	1,1%	1,9%	1,6%	1,7%	1,7%
Fundamental 1ª a 4ª série	12,5%	12,9%	18,0%	23,4%	13,2%	15,3%	15,7%
Fundamental 5ª a 8ª série	14,2%	16,4%	23,8%	20,1%	17,7%	17,9%	18,2%
Ensino Médio Incompleto	9,8%	8,6%	10,2%	8,1%	9,7%	7,7%	8,5%
Ensino Médio Completo	35,4%	29,5%	24,2%	25,4%	27,6%	28,3%	28,1%
Superior Incompleto	5,0%	5,1%	7,6%	3,6%	5,6%	5,8%	5,5%
Superior Completo	11,4%	15,0%	9,1%	10,1%	15,5%	17,0%	14,8%
Pós-graduação	7,2%	8,3%	4,3%	5,5%	6,8%	4,4%	5,5%
Não Sabe	2,4%	1,9%	1,6%	1,7%	2,2%	1,8%	1,9%
Total	100%	100%	100%	100%	100%	100%	100%
Observações	9.688	12.573	10.147	20.184	31.538	74.974	159.104

Fonte dos dados básicos: ENEM 2008

As regiões metropolitanas, por sua vez, possuem políticas educacionais e características sociais diferenciadas e que podem influenciar o desempenho de habitar certa localidade. Assim, foram incluídas seis variáveis dummies regionais, referentes às regiões metropolitanas de Recife (**reg_re**), Salvador (**reg_sa**), Porto Alegre (**reg_pa**), Rio de Janeiro (**reg_rj**), São Paulo (**reg_sp**) e Belo Horizonte (**reg_bh**).

QUADRO 3 - Estatísticas descritivas sobre gênero e cor/raça por região metropolitana

	Observações	Sexo		Cor/Raça	
		Homem	Mulher	Branco	Não-Branco
Total RM	159.104	38,9%	61,1%	55,3%	44,7%
Salvador	9.688	38,6%	61,4%	22,0%	78,0%
Recife	12.573	38,8%	61,2%	43,6%	56,4%
Porto Alegre	10.147	37,6%	62,4%	82,0%	18,0%
Belo Horizonte	20.184	37,9%	62,1%	39,5%	60,5%
Rio de Janeiro	31.538	37,2%	62,8%	49,4%	50,6%
São Paulo	74.974	40,1%	59,9%	64,7%	35,3%

Fonte dos dados básicos: Microdados do ENEM 2008

Por fim, temos as variáveis que identificam a rede administrativa da escola que o candidato frequentou durante o ensino médio. A rede Federal (**esc_fed**) compõe apenas 2% da amostra, o que demonstra seu tamanho pequeno em relação às

outras redes, as escolas estaduais (**esc_estad**) é a origem predominante na amostra, de cerca de 67% dos candidatos do exame, a rede privada (**esc_priv**) agrega 29% dos candidatos. A rede municipal (**esc_munic**) possui 2% da amostra e sua representação pequena se deve ao fato de existirem poucas escolas municipais que oferecem o ensino médio, a maioria das escolas dessa rede possui apenas o ensino fundamental disponível¹⁰.

4.2 Análise Descritiva das Variáveis Escolares utilizadas na Correção do Viés de Auto-Seleção

A principal dificuldade de implementar a correção de Heckman é a necessidade de se possuir informações sobre as pessoas as quais não possuímos a nota do exame. Contudo, as pessoas sem nota no exame são justamente aquelas que não forneceram quaisquer outras informações sociodemográficas, isto é, não sabemos nada sobre elas. Assim, foi necessário uma estratégia de identificar, pelo menos aproximadamente, esse grupo de pessoas ausentes no exame, com base em informações disponíveis de pessoas presentes na prova.

As variáveis utilizadas no primeiro estágio são todas referentes à escola em que o candidato está concluindo o ensino médio e devem permitir, em algum grau, uma separação dos estudantes que fizeram a prova, daqueles que teoricamente não a fizeram. A nossa hipótese é que a homogeneidade dos alunos de uma escola pode ser usada para determinar as características dos ausentes. Assim, os candidatos de uma escola com baixa participação no ENEM podem ser semelhantes aos seus colegas de turma que não compareceram ao exame.

Com respeito às características dos docentes das escolas, , a porcentagem destes com nível superior de ensino e com licenciatura é de quase 100%, de modo que tanto escolas pouco eficientes como as mais eficientes possuem quase

¹⁰ Segundo a Lei de Diretrizes e Bases da Educação Básica (LDB), os municípios devem incumbir-se de oferecer a educação infantil e o ensino fundamental, com a possibilidade de ofertar o ensino médio somente quando houver recursos suficientes e as necessidades da sua área de competência estiverem supridas BRASIL (1996).

a mesma proporção de docentes com licenciatura, o que torna esta variável pouco eficiente na discriminação dos estudantes.

A idade dos docentes, por sua vez, varia bastante entre instituições de ensino e também entre regiões metropolitanas, e pode ser uma influência importante, como motivação e preocupação, sobre todos os alunos de uma mesma escola. A idade mínima observada, considerando apenas as escolas que estão situadas nas regiões metropolitanas trabalhadas, é de 18 anos, e a idade máxima entre os docentes é de 80 anos. O quadro a seguir mostra as estatísticas descritivas da idade dos docentes, sem considerar o agrupamento por escola, por região metropolitana. Dentre as regiões metropolitanas, a de Belo Horizonte se destaca pela maior concentração de professores jovens.

Tabela 4 - Estatísticas descritivas da idade dos docentes por região metropolitana

	Obs.	Média	Desvio Padrão	Mínimo	Máximo
Recife	6.667	41,43	10,14	19	80
Salvador	5.469	41,72	9,51	18	77
Belo Horizonte	8.119	39,81	9,83	18	77
Rio de Janeiro	20.367	41,69	10,27	19	79
São Paulo	34.650	41,25	9,51	18	70
Porto Alegre	5.871	41,88	10,05	18	78
Total	81.143	41,31	9,85	18	80

Fonte dos dados básicos: Microdados do Censo Escolar 2008

Uma outra característica dos docentes que influencia na qualidade das aulas e, conseqüentemente, sobre a acumulação de capital humano dos alunos, é a sua qualificação profissional, que foi medida pela posse de um título de pós-graduação e reflete o seu domínio sobre os assuntos lecionados. Dentre os títulos de pós-graduação que os docentes declararam possuir, o de especialização foi o mais recorrente e, em seguida, o de mestrado e o de doutorado. A expectativa é que as escolas com maior número de professores qualificados transmitam a seus alunos mais conhecimento. Com isso, calculamos a razão entre o número de docentes com algumas pós-graduação (especialização, mestrado ou doutorado) sobre o total de docentes, por instituição de ensino.

Tabela 5 - Estatísticas descritivas da proporção de docentes com pós-graduação nas Escolas por região metropolitana

	N	Média	Desvio Padrão	Mínimo	Máximo
Recife	455	0,43	0,24	0	1
Salvador	302	0,48	0,24	0	1
Belo Horizonte	526	0,34	0,24	0	1
Rio de Janeiro	1.332	0,28	0,23	0	1
São Paulo	2.437	0,25	0,29	0	1
Porto Alegre	394	0,39	0,24	0	1
Total	5.446	0,31	0,27	0	1

Fonte dos dados básicos: Microdados do Censo Escolar 2008

Para Soares (2008), o papel da escola na educação dos estudantes não se limita em oferecer professores qualificados e bem motivados, como já foi discutido no segundo capítulo. A infraestrutura adequada permite que o docente consiga explorar melhor diversas maneiras para transmitir o conhecimento e o aluno, por sua vez, pode se sentir mais motivados em estudar em um ambiente agradável com recursos que facilitem o seu aprendizado.

Por outro lado, a infraestrutura direcionada para os recursos pedagógicos pode influenciar a maneira como os estudantes interagem com os professores, facilitando aquilo que é aprendido em sala de aula. Por isso, incluímos em nossa análise características de escola que incluem a existência de: laboratório de informática, laboratório de ciências, biblioteca, acesso à internet e banda larga disponível. A tabela a seguir resume a proporção de escolas que possuem cada um desses quesitos:

Tabela 6 - Proporção total das escolas por tipo de infraestrutura

Variável	Média
Possui Laboratório de Informática	81,00%
Possui Laboratório de Ciências	48,20%
Possui Biblioteca	85,70%
Possui Acesso à Internet	94,60%
Possui Banda Larga	74,60%

Fonte dos dados básicos: Microdados do Censo Escolar 2008

4.3 Análise Descritiva das Variáveis Utilizadas na Imputação da Proficiência Cognitiva do ENEM para a PME

Nesta seção, foram analisadas as variáveis comuns a ambas as bases de dados, que foram utilizadas em nosso modelo de imputação da nota do ENEM para os indivíduos da PME. O nosso objetivo com a descrição das variáveis é apresentar possíveis discrepâncias referentes à distribuição de características entre os indivíduos das duas amostras.

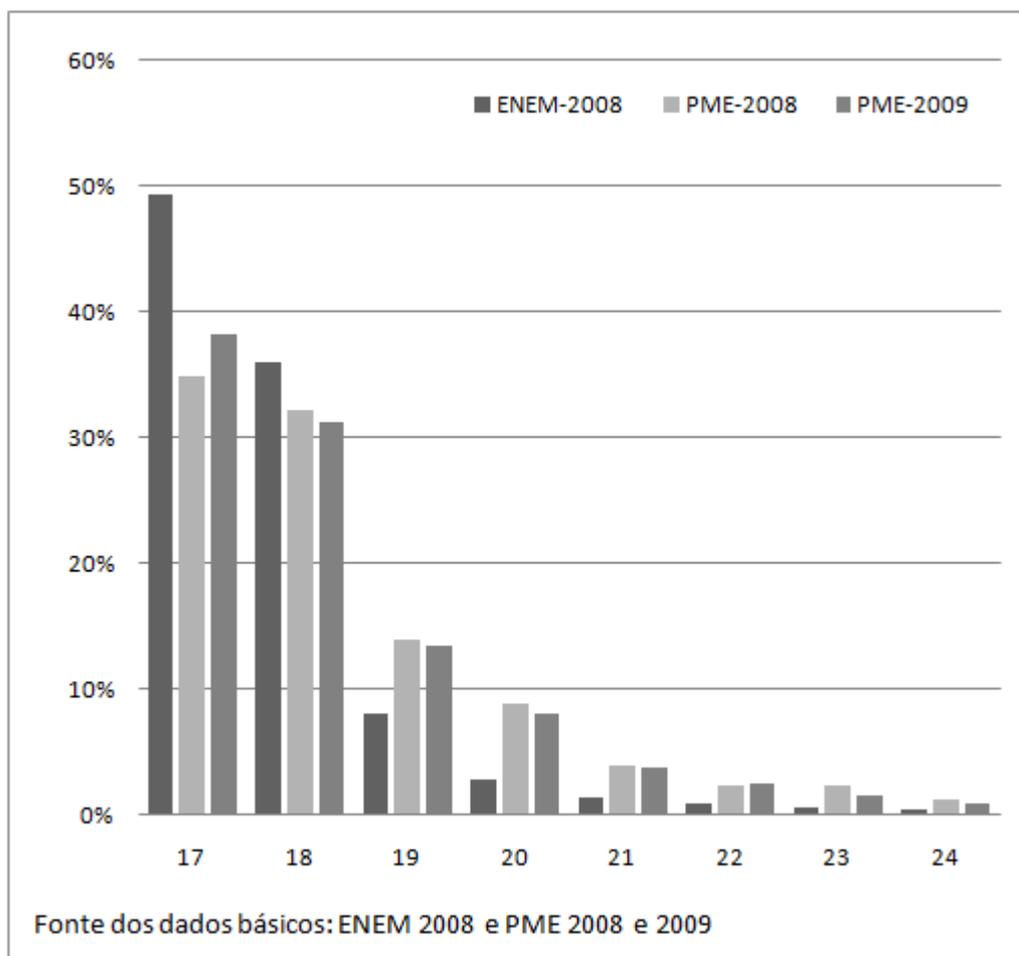
Em primeiro lugar, consideramos a idade do indivíduo. Na base de dados do ENEM de 2008, restringimos a amostra aos indivíduos que estavam cursando o último ano do ensino médio em 2008. Apesar disso, a idade dos indivíduos gira em torno dos 17 anos, idade considerada normal para o final do ensino básico.

O gráfico 4 mostra a comparação de idade entre as duas bases de dados, e aponta uma discrepância na distribuição dos indivíduos. Podemos observar claramente que existe um viés na amostra do ENEM, pois os indivíduos com idade entre 17 e 18 anos aparecem mais na base do ENEM do que na PME. Isso é, no entanto, natural, uma vez que, enquanto a PME é construída como uma amostra aleatória, a amostra do ENEM possui uma seleção positivamente correlacionada com essa faixa etária.

A diferença na distribuição de observações por idade não é muito grande, sendo assim, esperamos que ela não seja um problema no processo de imputação da

proficiência cognitiva na PME, desde que a distribuição condicional seja semelhante.

GRÁFICO 4 - Distribuição dos indivíduos por idade na base de dados do ENEM e PME

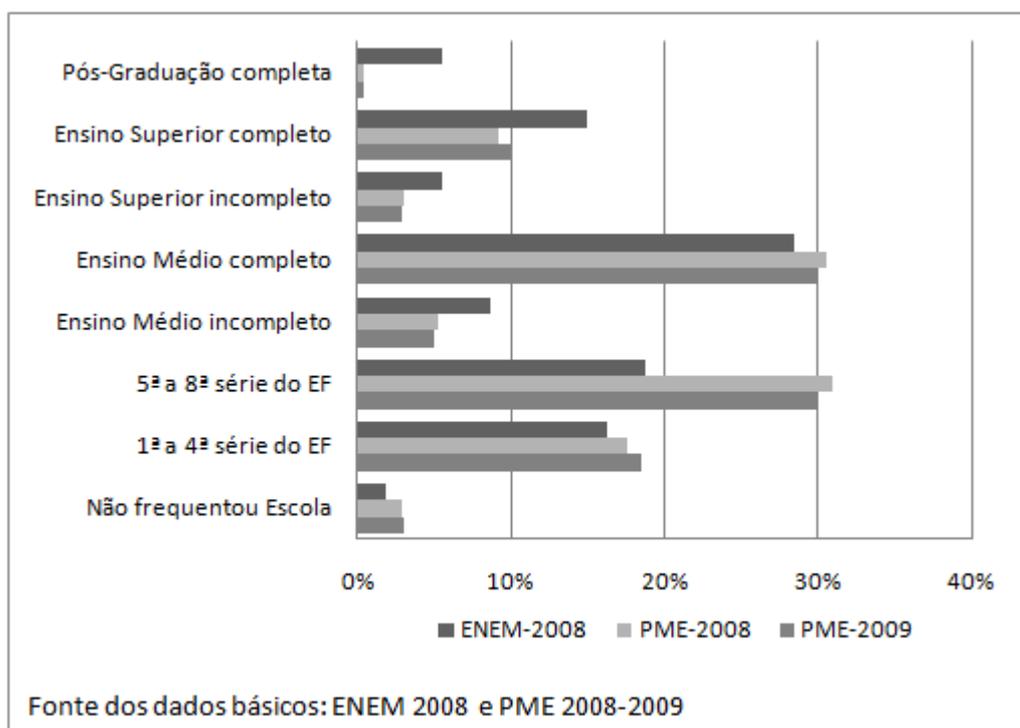


Em relação ao sexo dos indivíduos nas duas bases de dados, considerando o período analisado, a presença do sexo feminino é muito maior no ENEM (61,3%), do que na PME 2008 (50,2%) e na PME 2009 (55,1%). Com respeito a cor/raça dos indivíduos, existe uma proporção menor de pessoas que se declararam pretas ou pardas no ENEM (44,7%), do que na PME 2008 (51,7%) e na PME 2009 (49,1%).

Em seguida, comparamos as variáveis familiares que influenciam a proficiência do indivíduo, em especial, a escolaridade da mãe, a renda familiar e a quantidade de pessoas que moram no mesmo domicílio.

A escolaridade da mãe, na PME, foi medida pela escolaridade da pessoa de referência ou se cônjuge, sendo do sexo feminino. Com isso, foi possível construir a variável da maneira como está disponível no questionário do ENEM, para que pudéssemos utilizá-la na imputação. Os valores possíveis para escolaridade da mãe, assim como a distribuição percentual para os indivíduos das duas amostras podem ser visualizadas no gráfico N° 5.

GRÁFICO 5 - Distribuição dos indivíduos segundo escolaridade da mãe na base de dados do ENEM e PME



No gráfico acima, podemos destacar que, na base de dados do ENEM 2008, existe uma concentração de indivíduos cuja mãe possui um grau de escolaridade mais elevado, quando comparamos com os dados da PME. Isso pode ser uma fonte de viés de seleção, já que os indivíduos com mães mais escolarizadas estão mais propensos a comparecer ao exame do ENEM.

A renda familiar, por sua vez, destaca-se como uma das variáveis mais importantes para explicar a proficiência cognitiva dos jovens e também pode ser muito importante para a situação ocupacional dos mesmos. Por hipótese, ela pode garantir uma maior segurança aos jovens em suas decisões, permitindo-lhes arriscar mais sem sofrer grandes consequências econômicas. Além disso, pode

postergar a saída do jovem da casa dos pais e, com isso, também a entrada do mesmo no mercado de trabalho por necessidade (CARVALHO, 2009).

Sobre a construção da variável de renda familiar, no questionário socioeconômico do ENEM, obtivemos a renda agregada do domicílio, isto é, de todos que moram com o jovem, sem fazer nenhum tipo de distinção quanto à origem da renda, que pode ser fruto do trabalho, aluguéis, aplicações, pensões, etc. Por outro lado, o questionário da PME possui uma informação incompleta, referente à renda proveniente do trabalho.

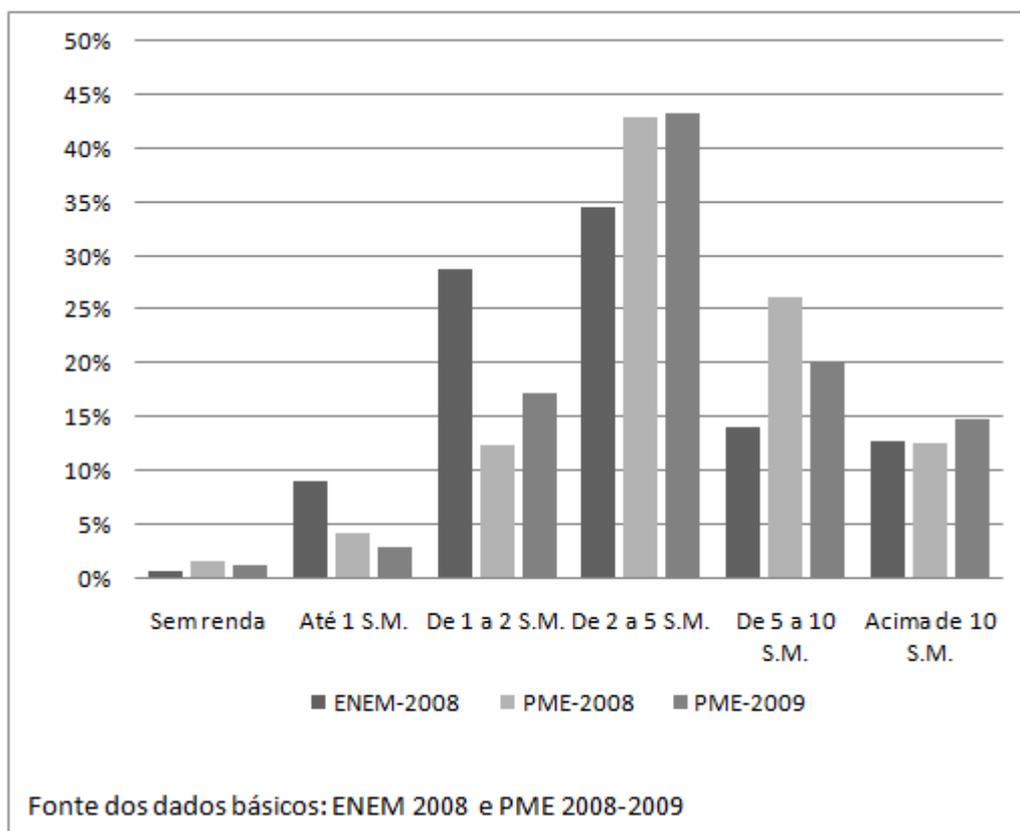
Para corrigir esse problema, utilizamos o método proposto por Machado & Ribas (2009), e imputamos a renda do não-trabalho na PME de 2008 e 2009, utilizando a PNAD do mesmo ano, cujas variáveis possibilitaram estimar a renda de todas as fontes¹¹. Após a imputação da renda do não-trabalho na PME, a renda foi deflacionada para agosto/2008¹² e separada em faixas, para equipará-la entre os indivíduos da PME e do ENEM.

O gráfico N° 6 mostra a distribuição percentual dos indivíduos segundo as faixas de renda e deixa claro que existe uma concentração maior de jovens de renda familiar mais baixa na base de dados do ENEM, quando comparamos com a PME. Novamente, os mecanismos de incentivo da prova do ENEM podem estar atuando de forma a atrair uma proporção maior de pessoas de baixa renda, como o uso da nota do exame como vestibular na concessão de bolsas a estudantes de baixa renda pelo ProUni (Programa Universidade para Todos).

¹¹ A densidade de kernel foi estimada para o logaritmo da renda total na PNAD e na PME, para os anos de 2008 e 2009, como forma de analisar o ajustamento dos dados. O gráfico da distribuição pode ser visualizado no gráfico A.1 em anexo.

¹² O indivíduo na base do ENEM respondeu o questionário com base em valores de agosto de 2008, quando o exame foi realizado. Os indivíduos na PME, por sua vez, responderam com base na época da pesquisa, isto é, em janeiro, maio e setembro de 2009. Por isso, utilizamos o deflator sugerido por Corseuil & Foquel (2002) para as pesquisas do IBGE, em especial, para a PME.

GRÁFICO 6 - Distribuição percentual dos indivíduos segundo faixa de renda familiar na base de dados da PME e ENEM

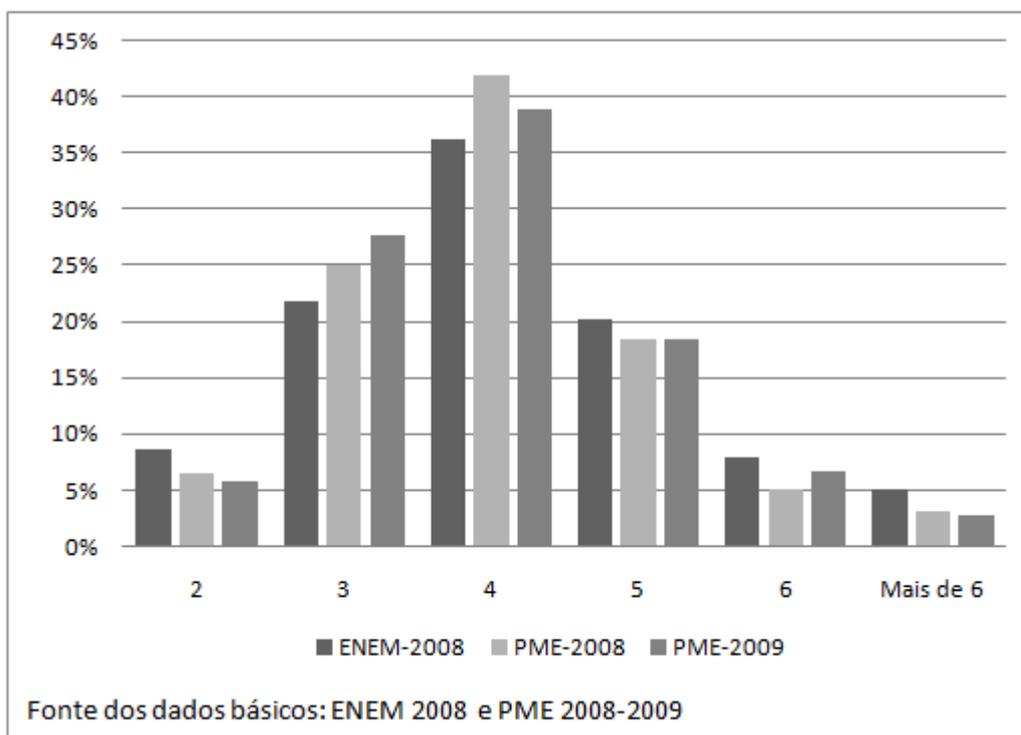


A última variável familiar considerada nas duas bases de dados é o número de pessoas que habitam o mesmo domicílio que o jovem. No questionário do ENEM, esse número vai de um, se a pessoa mora sozinha, até seis, se a pessoa mora com seis ou mais pessoas, ou seja, existe um truncamento neste valor.

Na PME, é possível contar quantas pessoas moram com os indivíduos, não existindo truncamento, mas, para efeito de igualdade das variáveis, optou-se por construí-la da mesma forma. O gráfico 7 sintetiza essa variável para ambas as bases de dados e mostra que a distribuição da variável indicativa do número de moradores por domicílio¹³ é bastante similar para as duas bases de dados.

¹³ Os dados excluem aqueles jovens que moram sozinhos. Isso aconteceu porque identificamos os jovens que estão na condição familiar de "filho" segundo o questionário da PME. Dessa forma, deve existir um chefe de família no domicílio, o que exclui a possibilidade do jovem morar sozinho. Para manter a homogeneidade das observações, não foram considerados tais casos na base do ENEM.

GRÁFICO 7 - Distribuição percentual dos indivíduos segundo a quantidade de moradores no domicílio na base de dados da PME e ENEM

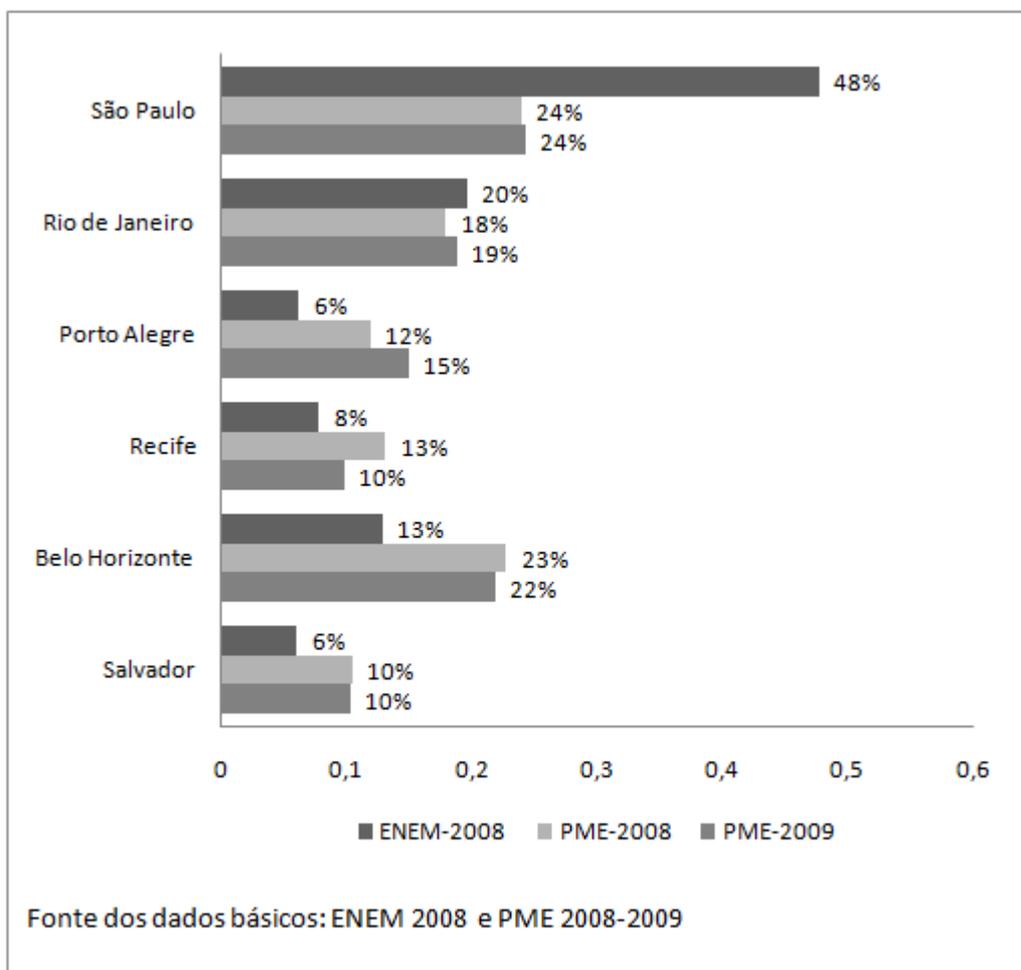


Por fim, analisamos a distribuição dos indivíduos segundo a região metropolitana em que residem. O gráfico Nº 8 mostra algumas divergências consideráveis na distribuição dos indivíduos entre as duas bases de dados. Existe uma concentração muito grande de indivíduos na região metropolitana de São Paulo, na base de dados do ENEM, comparativamente à PME.

A frequência de indivíduos na região metropolitana do Rio de Janeiro mostrou-se a mais estável entre as bases. Contudo, essa divergência na variável indicativa da região pode ocorrer como causa do desenho amostral da PME, que é feito com base em uma amostra aleatória de estratos na região metropolitana (municípios e pseudomunicípios), alguns setores dentro dos estratos e algumas unidades domiciliares dentre dos setores. Se considerarmos a expansão da amostra da PME pela variável que pondera pelo desenho amostral, então a divergência entre

a distribuição de indivíduos na PME e no ENEM, para as seis regiões metropolitanas, praticamente desaparece¹⁴.

GRÁFICO 8 - Distribuição percentual dos indivíduos segundo a região metropolitana em que mora



A análise descritiva das variáveis possibilitou observar que existe uma concentração de indivíduos, com determinadas características, em uma base de dados, que não é observada na outra. A causa dessas divergências pode ser a auto-seleção que ocorre na base de dados do ENEM.

Como já foi dito anteriormente, a PME tem característica aleatória, enquanto que o ENEM é a porta de entrada do jovem para o ensino superior, tanto por ser parte

¹⁴ A distribuição percentual dos indivíduos por região metropolitana utilizando a amostra expandida da PME, que pondera as observações com base no desenho amostral da pesquisa, pode ser conferida no gráfico anexo A.2

do vestibular de algumas instituições de ensino, como por ser utilizado para a concessão de bolsas de estudos, além de ser voluntário.

Essa diferença na distribuição de observações entre as bases de dados pode configurar em uma diferença na distribuição da nota do ENEM e da nota imputada na PME.

Ao observar os gráficos anteriores, podemos notar que existe, de fato, um viés de seleção na amostra do ENEM, quando comparamos com os dados da PME. No capítulo anterior, utilizamos o modelo de correção de Heckman para testar essa hipótese, a qual comprovamos como afirmativa, dispondo de certas hipóteses para a estimação do modelo.

Ao contrário da base de dados do ENEM, que permite vincular informações do Censo Escolar, o qual nos baseamos para construir a base de dados do primeiro estágio de Heckman, essa possibilidade inexistente na PME. Dessa forma, optando pela parcimônia na aplicação dos métodos, apontamos a existência do viés de seleção do ENEM, que pode enviesar as nossas estimações. Contudo, a análise gráfica mostra que a diferença não é demasiadamente grande, e a inclusão de uma correção do viés de seleção demandaria outras hipóteses para a sua viabilidade na PME, diferentes daquelas feitas no terceiro capítulo.

4.4 Análise Descritiva da Variável de Situação Ocupacional na Base de Dados da PME

O objetivo de investigar como a proficiência influencia a situação ocupacional dos jovens, doze meses após a conclusão do ensino médio, é feita por meio da variável que indica a frequência do jovem em uma instituição de ensino e de outra que indica se ele está ativo no mercado de trabalho. A variável de situação ocupacional, relembrando, é formada por quatro categorias, são elas: não trabalha e não estuda (1), trabalha e não estuda (2), trabalha e estuda (3), não trabalha e estuda (4).

Uma leitura mais clara da variável dependente é feita em termos percentuais, como em uma matriz de transição. No quadro Nº 4, as linhas se referem à

situação inicial em 2008 ou 2009, e as colunas, por sua vez, se referem à situação final em 2009 ou 2010. A diferença observada entre 2008 e 2009 é pequena, como se pode observar no quadro abaixo.

Quadro 4: Matriz de transição da situação ocupacional em doze meses

			Situação final				Total
			Só estuda	Trabalha e estuda	Trabalha e não estuda	Não trabalha e não estuda	
Situação inicial	2008	Só estuda	33%	11%	31%	26%	100%
		Trabalha e estuda	8%	19%	58%	15%	100%
		Total	23%	14%	42%	22%	100%
	2009	Só estuda	30%	9%	33%	28%	100%
		Trabalha e estuda	8%	22%	56%	14%	100%
		Total	22%	14%	41%	23%	100%

Fonte dos dados básicos: IBGE - PME 2008, 2009 e 2010

Na situação de quem estuda e trabalha ao mesmo tempo, observamos uma tendência mais acentuada dos jovens não continuarem os estudos e se dedicarem inteiramente ao trabalho (58% em 2008 e 56% em 2009). Outro ponto interessante é o acentuado número de jovens que tornam-se inativos e param de estudar. Ao considerar a soma da frequência das duas situações em que o jovem não estuda, 73% e 70% parou de estudar após a conclusão do ensino médio, respectivamente, em 2008 e 2009.

Na situação de quem apenas estuda, os dados mostram um cenário similarmente adverso. A soma da frequência relativa das duas situações finais, em que o jovem não estuda, teremos 57% e 61%, respectivamente, para quem concluiu o ensino médio em 2008 e 2009. Ou seja, dos jovens que estão terminando o ciclo básico educacional, mais da metade interrompeu os estudos.

Esse fenômeno pode ser reflexo do custo direto e indireto de aumentar a escolaridade. A existência de restrições orçamentárias e/ou individuais, podem fazer com que o jovem deixe de estudar. Além disso, mostra que uma parcela considerável da população não está acumulando capital humano, o que pode influenciar na oferta de mão-de-obra qualificada no mercado de trabalho.

5. FUNÇÃO DE PRODUÇÃO EDUCACIONAL

A função de produção educacional estimada tem, por objetivo, identificar os fatores associados à proficiência cognitiva dos jovens concluintes do ensino médio e que possuíam entre 15 e 24 anos em 2008, considerando características individuais, familiares e escolares. A finalidade deste capítulo é apresentar ao leitor os resultados dos modelos econométricos utilizados para estimar a função de produção educacional.

O primeiro modelo de regressão apresentado é estimado utilizando o método de mínimos quadrados ordinários (MQO) para explicar a nota da prova objetiva e, posteriormente, a nota da redação, com variáveis explicativas pessoais, familiares e escolares, sendo que o resultado obtido condiz com o encontrado em trabalhos com o mesmo interesse. Posteriormente, são apresentados os resultados obtidos pela estimação do modelo de correção de Heckman para o viés de auto-seleção e pelos modelos hierárquicos lineares, utilizados para corrigir a quebra de hipótese das amostras independentes e identicamente distribuídas.

5.1 Estimação da Função de Produção Educacional por MQO

Os resultados da regressão linear por MQO, para a prova objetiva, podem ser analisados na Tabela Nº 7, onde a nota desta prova do ENEM 2008 é a variável dependente e as variáveis pessoais, familiares e escolares são independentes. É importante ressaltar que, salvo a variável idade, que foi centralizada, todas as demais são variáveis dicotômicas e seus coeficientes devem ser interpretados como diferenças em relação à categoria de referência.

A variável indicadora da idade do candidato foi centralizada em 17 anos, considerada normal para o final do ensino médio, e seu coeficiente sobre a proficiência é negativo e estatisticamente significativo. Por isso, segundo o modelo estimado, os candidatos possuem 1,33 ponto a menos que a média, na prova objetiva, para cada ano adicional superior a 17 anos.

Esse efeito parcial negativo pode ser associado ao atraso escolar, causado por alguma interrupção intermitente ou repetência, ou candidatos que terminaram o ensino médio a mais tempo e voltaram a estudar. O avanço etário, relativamente à série cursada, parece ser um obstáculo ao aprendizado, refletido no baixo aproveitamento do aluno, mesmo sabendo que a relação entre idade e proficiência possa ser simultânea. Neste caso, a causa da idade avançada pode ser justamente o baixo rendimento escolar. O mesmo resultado foi encontrado por Sampaio (2009) com dados do ENEM 2005 e 2006, e também por Golgher (2010) para o ENEM 2005.

A escolaridade da mãe se mostrou um fator determinante para a proficiência escolar dos filhos, uma vez que a aquisição de capital humano na família pode depender, em parte, da escolaridade dos pais. O modelo estimado deixa claro que a proficiência dos filhos cresce de forma não linear conforme a escolaridade da mãe aumenta, quando comparamos com aqueles cuja mãe nunca frequentou a escola, que é a categoria de referência. A exceção é o candidato que desconhece a escolaridade da mãe, para o qual o efeito não difere estatisticamente da categoria base.

O efeito da renda familiar (apenas de quem mora no mesmo domicílio) sobre a proficiência, é positivo e não linear com respeito às diferentes faixas de renda, mesmo controlando pela escolaridade da mãe, que, por hipótese, é positivamente correlacionada com a renda permanente da família. Essa observação confirma a relação positiva entre o nível socioeconômico e o desenvolvimento educacional das pessoas discutida no segundo capítulo.

No entanto, afirmou-se anteriormente que morar com irmãos acarreta um menor aproveitamento escolar pelo fato da divisão dos recursos destinados a sua educação. O fato é que o tamanho da família no domicílio diminui a quantidade total de recursos financeiros disponíveis para a educação de uma maneira geral, de modo que este fique abaixo do patamar ideal. Para tentar captar o efeito da quantidade de moradores no domicílio do candidato sobre a distribuição de renda intrafamiliar, foram calculadas as interações de cada faixa de renda com a quantidade de moradores no domicílio.

A variável “moradoresXrenda1sm”, por exemplo, refere-se à interação entre a quantidade de moradores no domicílio cuja renda familiar é de até 1 salário mínimo, e o seu coeficiente negativo deixa evidente que a vantagem dessa renda familiar sobre a proficiência, em relação à categoria de referência (nenhuma renda familiar), decresce conforme o aumento na quantidade de moradores no domicílio.

O mesmo efeito acontece para famílias com faixas de renda mais elevadas, exceto quando a renda excede 10 salários mínimos, para o qual o coeficiente da variável moradoresXrenda10sm não é estatisticamente diferente de zero. Este resultado pode sugerir que, para níveis de renda bastante elevados, a demanda educacional dos entes mais jovens da família é suprida financeiramente em sua totalidade, não sendo limitada por uma restrição orçamentária da família e as suas decisões de alocação de recursos entre os membros.

Quanto ao sexo do candidato, observamos que a nota média da prova objetiva dos homens é superior à das mulheres, indicado pelo coeficiente negativo e significativo para o sexo feminino na regressão linear por MQO. Em contrapartida, o sexo feminino obteve melhor resultado na prova de redação, como fica evidente na Tabela Nº 8, que mostra os resultados obtidos utilizando a nota da redação como variável dependente. Esse resultado complementa a discussão do segundo capítulo, sobre a diferença de desempenho entre sexos, tendo os homens maior vantagem em questões lógicas, e as mulheres maior vantagem em provas argumentativas.

Os indivíduos pretos/pardos demonstraram menor nível de proficiência do que brancos/amarelos, e aponta a sua maior desvantagem em termos de sucesso escolar. Embora a sua causa possa vir da maior concentração de candidatos de cor preta/parda nas faixas de renda familiar mais baixas, estamos controlando pelo nível de escolaridade da mãe e pela renda familiar, de modo que existem outros fatores, ligados à cor/raça, que explicam essa desigualdade étnica sobre a proficiência educacional dos jovens.

A variável indicativa de trabalho do jovem registra um efeito negativo e significativo, que pode ser uma consequência do menor tempo disponível

dedicado à educação, reduzindo o aproveitamento escolar. Ressaltando que estamos controlando a proficiência pela renda familiar do candidato, de tal forma que essas variáveis captam o efeito parcial do nível socioeconômico sobre a educação. No entanto, como a variável de renda domiciliar é categorizada, pode existir uma certa correlação dentro de cada categoria de renda com essa variável.

A frequência a cursos preparatórios aumenta diretamente o aproveitamento no exame, uma vez que eles são organizados com esse objetivo. Tais cursos podem corrigir o desempenho escolar pífio de um aluno que obteve educação prévia de baixa qualidade, mas como os jovens que frequentam uma escola de baixa qualidade são também aqueles que não podem pagar por um ensino de qualidade, também é pouco provável que possuam recursos financeiros para adquirir este serviço.

A variável “noturno” apresenta um coeficiente negativo, indicando uma desvantagem para aqueles que estudam no período noturno, em relação aos que estudam no período diurno, mesmo controlando pela rede de ensino. Apesar de uma parte considerável dos candidatos que afirmam estudar no período noturno não trabalhar, tais indivíduos podem alocar o seu tempo em outras atividades, como trabalho informal ou doméstico, o que pode revelar a menor prioridade à educação.

Os resultados mostram que existem diferenças entre as redes de ensino sobre a proficiência média de seus alunos. O sinal positivo dos coeficientes da rede municipal, federal e privada mostra que todas são melhores, em termos de nota no ENEM, do que a rede estadual (categoria de referência). Além disso, podemos notar que, apesar da diferença entre a rede municipal e a estadual, os candidatos que estudaram nessas redes possuem, em média, uma proficiência abaixo daquela obtida por alunos provenientes da rede particular ou federal, um sinal da qualidade do ensino oferecido por cada uma destas últimas.

A rede particular de ensino diferencia-se por sua maior autonomia e agilidade na implementação de políticas educativas, e por deterem incentivo maior para atingir taxas de aprovação em vestibulares do que as redes estaduais e municipais, como consequência da concorrência do mercado.

A rede federal se destaca em relação às demais, contudo, a vantagem dos alunos dessa rede sobre o desempenho no exame pode não ser um reflexo direto da escola. Em geral, a rede federal possui processos seletivos para admissão de alunos, em que apenas os melhores colocados conseguem se matricular. Isso pode ser um indicativo de fatores não-observáveis, uma vez que controlamos por outros fatores pessoais e familiares.

Os resultados não rejeitam a hipótese de que haja uma alocação de recursos intrafamiliares disponíveis para a educação, seja financeiro ou até mesmo a atenção dos pais, entre filhos. Esse fenômeno foi medido pela variável indicativa de morar com irmãos, e que apresenta um coeficiente negativo, justificando o disposto no segundo capítulo, que o menor aproveitamento escolar de jovens que moram com irmãos pode ser visto como uma consequência da divisão dos recursos educacionais entre eles.

Por fim, foram observadas diferenças entre as regiões metropolitanas, por diversas razões. De acordo com o modelo contextual apresentado na Figura Nº 1, as escolas de regiões diferentes podem estar sujeitas a políticas educacionais, valores socioculturais e religiosos, legislação diferenciada e demandas sociais diferentes, o que influencia a quantidade de recursos que as escolas públicas recebem e as diretrizes de gestão desses recursos por cada escola.

A região metropolitana de Belo Horizonte apresentou resultado médio superior ao das demais regiões metropolitanas, como pode ser visto pelo coeficiente negativo de todas as variáveis regionais, tomando a região metropolitana de Belo Horizonte como referência. Um resultado similar resultado é obtido por Golgher (2010a), que introduz uma variável *dummy* para cada UF do Brasil para estimar a nota do ENEM 2005, obtendo coeficientes negativos para todos os Estados, tomando Minas Gerais como UF de referência.

Com respeito aos determinantes da nota de redação estimados por MQO, as regiões metropolitanas de Salvador e Porto Alegre mostraram-se superiores à de Belo Horizonte, em termos de proficiência média, como pode ser visto na Tabela Nº 8, diferente do ocorrido quando analisamos a prova objetiva.

Os resultados da regressão linear por MQO para a nota de redação apontam algumas diferenças entre esta e a prova objetiva. Já foi dito anteriormente que o sexo feminino se saiu melhor na prova de redação, enquanto o sexo masculino obteve melhor resultado na prova objetiva. Além disso, os resultados obtidos para a prova de redação mostram que os fatores a ela associados possuem menor impacto sobre a sua nota, quando comparados com o impacto que estas mesmas variáveis possuem sobre a nota da prova objetiva.

A escolaridade da mãe e a renda familiar, importantes para explicar a nota da prova objetiva, possuem um efeito relativamente menor sobre a proficiência medida pela redação. O papel dos fatores não-observados do candidato parece ser maior quando ele possui liberdade para expressar o seu pensamento.

A variável indicadora se o candidato fez cursinho mostrou-se relevante tanto para a parte objetiva, quanto para a discursiva, mas o seu coeficiente para a redação é relativamente menor. Os cursos preparatórios possuem uma curta duração, geralmente inferior a um ano, e é possível que eles consigam treinar os candidatos para responder a questões objetivas, mas não consigam ensiná-los a desenvolver um raciocínio lógico discursivo.

Isso pode significar que o desenvolvimento cognitivo, como requerido pela prova de redação, pode ser resultado de uma educação que se desenvolve na pessoa a médio ou longo prazo. O baixo poder explicativo das variáveis independentes para a nota de redação, mesmo controlando por atributos pessoais, familiares e escolares, pode indicar que a competência de expressar as ideias depende fortemente de fatores não observáveis.

5.2 Função de Produção Educacional com Correção para o Viés de Auto-Seleção do ENEM

O ENEM é um exame voluntário e, no caso desse ano, como já dito anteriormente, apenas a metade dos candidatos compareceu no dia da prova, há um problema de auto-seleção que surge pela não aleatoriedade na ausência do exame. Se houver um viés no sentido de que aqueles cujo rendimento escolar é mais elevado sejam mais propensos a fazer o exame, então deixamos de

observar o resultado de jovens menos preparados e, dessa forma, podemos subestimar os fatores que determinam a proficiência.

Por meio do Censo Escolar 2008, foi possível calcular a quantidade de alunos do 3º ano do ensino médio em cada escola do país e, anexando esse dado com a quantidade de candidatos que fizeram a prova do ENEM em 2008, calculamos o percentual de adesão de cada escola, referente a essa população de alunos. O resultado de uma análise de correlação da nota e o percentual de adesão mostra que existe uma relação positiva entre essas variáveis.

Ao ignorar esse viés de seleção da amostra, a regressão por MQO produz coeficientes que podem enviesar os reais efeitos dos determinantes da proficiência escolar, caso exista uma auto-seleção do tipo não-aleatória. Para corrigir esse viés, foi utilizado o modelo de Heckman em dois estágios.

No primeiro estágio, calculou-se a probabilidade de um indivíduo comparecer ao exame, com base em características da rede de ensino que frequentou. No segundo estágio, a proficiência cognitiva foi estimada com base nas mesmas variáveis explicativas utilizadas no modelo de regressão linear por MQO, mas agora o erro foi corrigido com a inclusão da razão inversa de Mills, obtida do primeiro estágio.

O fato do banco de dados não possuir nenhuma característica dos candidatos que não prestaram o exame demanda uma estratégia de identificação desse grupo para que a correção de Heckman possa ser feita.

Portanto, com base na presença relativa dos alunos do 3º ano do ensino médio de cada escola, selecionamos aquelas escolas cuja presença foi inferior a 10%, e assumimos que todos os estudantes dessas escolas faltaram ao exame, depois selecionamos as escolas com índice de participação superior a 90%, assumindo que os alunos dessas escolas compareceram ao exame.

A estratégia mostrou-se consoante ao fato de que as escolas com menor participação no exame também são aquelas cujos alunos possuem desempenho abaixo da média. Por fim, das 159.104 observações válidas utilizadas pelo modelo MQO, foram considerados como ausentes 5.521 candidatos, e como

presentes 2.081 candidatos, totalizando 7.602 observações usadas no primeiro estágio da correção de Heckman.

A correção do viés de seleção para a prova objetiva torna a leitura dos coeficientes menos distorcida e sublinha algumas mudanças em relação aos resultados obtidos via MQO em cujo viés está presente. Em primeiro lugar, o modelo de Heckman aponta que a correlação do erro do primeiro com o segundo estágio é positiva, indicando que os fatores não-observáveis que definem a propensão do indivíduo em comparecer ao exame são positivamente correlacionados com os fatores não-observáveis que explicam sua proficiência no exame.

Ao analisar o segundo estágio do modelo de Heckman na Tabela Nº 7, corrigido o viés pelo primeiro estágio, o efeito da idade sobre a nota é ligeiramente maior que no modelo MQO e o efeito da cor/raça permanece praticamente inalterado. Isso pode indicar que estes coeficientes não estariam sendo enviesados ao considerar o problema da auto-seleção.

A variável indicadora de residir com irmãos permanece significativa, quando comparada ao coeficiente obtido por MQO, sugerindo que a alocação de recursos entre irmãos explica a proficiência, mesmo consideramos uma amostra sem viés. Entretanto, a significância desta variável para a amostra auto-selecionada pode indicar que a alocação talvez seja relevante para a parcela dos candidatos mais capacitada.

O efeito da escolaridade da mãe no modelo de Heckman diminui para todos os níveis educacionais sobre a proficiência do filho, em relação ao modelo MQO. Ao considerar a amostra sem viés, observamos que a escolaridade da mãe estava sendo superestimada, não deixando de apontar que a vida educacional dos filhos é muito influenciada pela educação da mãe. Mais uma vez, destaca-se o papel da escolaridade dos pais na formação de capital humano na família.

O efeito de cada rede administrativa da escola também muda, pois agora frequentar uma escola federal ou privada garante um acréscimo maior na nota do candidato em relação aos estudantes da rede estadual, quando comparamos aos modelos anteriores. A rede municipal apresentou um coeficiente menor do que no

modelo MQO, isto é, ele estava sendo superestimado. Isso pode estar associado a maior chance de insucesso dos alunos de escolas públicas estaduais em relação aos oriundos das redes privada, federal, ou mesmo municipal.

As diferenças regionais tornaram-se menos acentuadas quando comparadas à região metropolitana de Belo Horizonte. A menor diferença também corresponde ao fato de que o viés é menor na região de Belo Horizonte, por agrupar o menor número de jovens que estudam em escolas cuja participação no exame não excede 5% da turma. Em outras palavras, a presença do estudante da região metropolitana de Belo Horizonte no ENEM 2008 foi mais efetiva.

Na estimação do modelo de Heckman para a nota de redação utilizamos o mesmo método para criação artificial da subamostra necessária ao primeiro estágio, que estima a propensão do indivíduo em fazer a redação. No entanto, somamos a esta subamostra o grupo de candidatos que fizeram a prova objetiva, mas não fizeram a redação.

Um dos principais resultados desta regressão é que a escolaridade da mãe como determinante da nota da redação é menor que no modelo MQO, exceto quando a mãe possui ensino superior completo ou pós-graduação. A variável “cursinho” também apresenta um coeficiente maior, mostrando que o modelo MQO estava subestimando tais coeficientes.

O coeficiente estimado da razão inversa de Mills, utilizada no segundo estágio para a correção do viés de auto-seleção, apresentou um coeficiente positivo, bem como o coeficiente de correlação (ρ) entre os erros do primeiro e do segundo estágio. Em outras palavras, os fatores não-observáveis que explicam a participação na prova de redação, são positivamente correlacionados com os fatores não-observáveis que explicam o rendimento do candidato na redação.

5.3 Estimação da Função de Produção Educacional por meio de Modelos Hierárquicos Lineares

A outra característica das bases de dados de educação que envia os coeficientes da regressão linear clássica é a sua estrutura hierárquica. As

observações individuais, em sua maioria, não são totalmente independentes, como postula uma das hipóteses básicas do método de mínimos quadrados ordinários. A consequência disso é que a correlação média (ou intra-classe) entre variáveis, considerando estudantes de uma mesma escola, é maior que a correlação média entre variáveis considerando indivíduos de escolas diferentes. A violação da hipótese de independência dos dados faz com que as estimativas de erros-padrão dos testes estatísticos sejam muito pequenos, o que pode resultar em resultados significativos, porém espúrios. (HOX, 2002, pág.05)

O primeiro passo foi estimar o modelo nulo HLM0, como discutido no terceiro capítulo, que tem como objetivo estimar as variâncias relativas a cada nível, aluno e escola, as quais podem ser analisadas pelo índice de correlação intra-classe (ICC). O ICC estimado indica que 40,59% da variância total da nota é explicada por características do nível da escola e 59,41% é explicada por fatores pessoais/familiares de cada candidato.

O modelo nulo utilizado para estimar a nota de redação como variável dependente mostra um resultado um pouco diferente. O índice de correlação intra-classe é de cerca de 13%, isto é, a variância da nota de redação explicada pelos fatores escolares são baixos, quando comparamos com o resultado da prova objetiva.

Ao analisar o SIMAVE, Soares (2005) estima um modelo com três níveis (alunos, turmas e escolas), e seu modelo nulo apresenta ICC de 12% em relação à variação devida à escola, para alunos da 4ª série. Por outro lado, Machado et. al. (2008) estimam um modelo hierárquico com três níveis (aluno, escola e município) utilizando a mesma base de dados, e encontra um ICC de 12,5% para a escola, quando considerados os alunos do 3º ano do ensino médio.

Em trabalhos que buscam o mesmo objetivo, mas por meio do Saeb, destacam-se os resultados obtidos por Albernaz et. Al. (2003), cujo modelo nulo para explicar a nota da prova de matemática do Saeb 1999 apresentou ICC de 28%. Para César & Soares (2001), estimam o ICC em 35,1%, considerando um modelo nulo em três níveis (alunos, escolas e UF). No entanto, como estes trabalhos

citados utilizam bases de dados diferentes e níveis diferentes de hierarquia, eles não podem ser diretamente comparáveis.

Um trabalho que se assemelha bastante ao nosso, por estudar alunos do 3º ano do ensino médio, por meio do Saeb 2001, e por utilizar um modelo hierárquico com a mesma estrutura que a nossa, e o de Andrade & Laros (2007), cujo ICC estimado no modelo nulo foi de 46%.

Em seguida, foi estimado outro modelo hierárquico, para a prova objetiva (modelo HLM1 na Tabela Nº 7) e para a redação (modelo HLM1r na tabela Nº 8), contendo as variáveis de primeiro nível, referentes ao candidato. O resultado indica uma diferença em relação aos coeficientes estimados pelos modelos de MQO e Heckman.

No modelo HLM1, para a prova objetiva, a inclusão das variáveis de primeiro nível faz com que o índice de correlação intra-classe (ICC) diminua em relação ao modelo nulo, com uma queda aproximada de dezesseis pontos percentuais. Mesmo assim, cerca de 24% da variância continua a ser atribuída a fatores de segundo nível, isto é, mesmo controlando fatores individuais, a escola ainda faz diferença na proficiência média de seus alunos. Como exposto no segundo capítulo, a escola possui uma série de atributos que influencia seus alunos, como a gestão dos recursos, programa pedagógico, etc. Tais fatores podem contribuir para a diferença de proficiência entre escolas, mesmo que estas sejam similares em alguns aspectos.

A inclusão das variáveis de primeiro nível para explicar a nota da redação, no modelo HLM1r da Tabela 8, ajuda a explicar parte da variância devida originalmente à escola, e esta se reduz em aproximadamente 23%. Quando comparamos este resultado com os modelos estimados para a prova objetiva, observamos que a escola parece ter um papel maior sobre o sucesso em exames padronizados com questões de múltipla escolha, do que na redação, onde o candidato é livre para expressar seu raciocínio. No resultado obtido por Andrade & Laros (2007), para os alunos do 3º ano do ensino médio na prova de matemática e língua portuguesa do Saeb 2001, após a inclusão das variáveis de primeiro nível, o ICC encontrado foi de 17%.

No terceiro modelo hierárquico (HLM2 na tabela 7 e HLM2r na tabela 8), foram incluídas, além das variáveis de primeiro nível, variáveis do nível da escola, como a rede administrativa e a região metropolitana em que ela se situa. A inclusão dessas variáveis tem por objetivo explicar parte da variância referente à escola e que não foi explicada pelas variáveis de primeiro nível.

A escolaridade da mãe possui menos influência quando consideramos o efeito da escola sobre seus alunos. É possível afirmar que existe uma concentração de alunos, cujo nível socioeconômico está acima da média, em escolas privadas, e alunos de famílias menos favorecidas em escolas públicas estaduais/municipais. Dessa maneira, o estoque de capital econômico/humano dos pais pode determinar a escolha da escola frequentada pelos filhos, bem como essa pode ser uma causa pela qual as escolas que apresentam resultados mais próximos aos esperados são mais disputadas pelas famílias mais ricas.

Em relação à influência da renda familiar sobre a proficiência, vemos que o efeito da família estar entre as duas faixas de renda mais elevadas se reduz, em comparação ao modelo de regressão linear e de Heckman. Isso pode sinalizar que a escola substitui, em algum nível, o benefício do capital econômico sobre o aprendizado dos jovens, por diversos motivos.

Neste caso, a escola pode oferecer um conhecimento que outrora seria possível somente por meio do investimento direto da família, que por sua vez, depende da renda disponível. Quando observamos a interação entre renda familiar e a quantidade de moradores no domicílio, notamos que a divisão de recursos intra-domicílio parece influenciar negativamente a proficiência, assim como nos outros modelos.

As famílias de classe social mais elevada possuem recursos suficientes para atender à demanda educacional dos filhos até o nível desejado. O modelo hierárquico também atribui maior proficiência aos candidatos que pertencem a famílias mais abastadas, bem como foi estimado pelas regressões anteriores utilizando MQO e Heckman.

A diferença entre a rede estadual e municipal de ensino diminui, quando utilizado o modelo hierárquico. A variável “**esc_munic**”, indicadora de o aluno frequentar a

rede municipal de ensino, deixa de ser positiva e passa a ser estatisticamente não significativa, sugerindo que a qualidade desta rede se equipara com a da rede estadual de ensino. Como explicado no terceiro capítulo, a desconsideração da estrutura hierárquica dos dados pode acarretar em coeficientes erroneamente significativos e diferentes de zero.

Em relação às diferenças entre regiões metropolitanas, a de Belo Horizonte continua sendo a melhor em termos de proficiência média, mesmo controlando a proficiência por outras variáveis. A diferença das outras em relação a esta mudou quando comparamos os coeficientes do modelo hierárquico com o MQO ou Heckman. Algumas regiões metropolitanas apresentam uma diferença de resultados, como São Paulo e Porto Alegre, cujos efeitos são menores do que em outros modelos, sinal de que a qualidade das escolas destas regiões metropolitanas não se distancia tanto das escolas da região metropolitana de Belo Horizonte.

A inclusão das variáveis de segundo nível contribui para explicar a heterogeneidade entre escolas. O ICC do modelo HLM2 apresentou uma diferença negativa de cerca de dez pontos percentuais em relação ao modelo anterior. Assim, pouco mais de 13% da variância da nota se deve a fatores escolares, enquanto que 87% da variância é explicada por atributos do primeiro nível. Este resultado é similar ao encontrado por Andrade & Laros (2007) para o Saeb 2001, no qual o modelo hierárquico com dois níveis (alunos e escolas) resultou em um ICC de 13,1%, depois de consideradas as variáveis de primeiro e segundo nível no modelo.

A estimação do modelo hierárquico, cujos resultados estão disponíveis na Tabela 8, considera as variáveis de primeiro e segundo nível para a prova de redação (HLM2r). O seu resultado mostra que, diferente da prova objetiva, a região metropolitana de Salvador não é estatisticamente diferente da de Belo Horizonte. Do mesmo modo, a região metropolitana de Porto Alegre, que possui um rendimento médio inferior, em relação à categoria de referência, na prova objetiva, possui uma nota média superior na prova de redação. Isso mostra que as escolas da RM de Porto Alegre preparam melhor seus estudantes com

respeito ao conjunto de competência de compreensão escrita, argumentação e utilização do conhecimento prévio.

Ao comparar o sinal das variáveis de todos os modelos estimados, podemos concluir que, apesar da correção dos vieses encontrados, a maioria das variáveis possui o mesmo sinal, independente do método econométrico utilizado.

O sinal dos coeficientes que são significativos, ao nível de 5%, para explicar a nota da prova objetiva e da redação, diverge quanto ao gênero dos candidatos, pois o sexo feminino obteve melhor desempenho na redação em relação ao sexo masculino, mas o mesmo não ocorre na prova objetiva.

A região metropolitana de Porto Alegre apresentou uma nota média menor que Belo Horizonte na prova objetiva, ocorrendo o contrário na prova de redação. A região metropolitana de Salvador também apresentou sinal divergente, com efeito positivo para a prova de redação, e negativo para a prova objetiva. No entanto, ela não é estatisticamente significativa para o modelo hierárquico.

Dessa forma, nossas estimações comprovam a importância de certas variáveis para explicar a proficiência cognitiva dos jovens, mesmo na ausência de vieses, que poderiam ocultar ou distorcer a importância de cada uma delas. Assim como encontrado pelos trabalhos citados ao longo do capítulo, os fatores socioeconômicos possuem grande influência sobre o aproveitamento escolar e aquisição de capital humano por parte dos jovens em idade escolar. Com relação ao efeito estimado dos fatores familiares, a escolaridade da mãe e a renda familiar, apesar de serem positivamente correlacionadas, possuem efeitos que se sobrepõe aos demais em questão de proficiência dos filhos.

Ao considerar o efeito das escolas, posto o que já foi dito acerca da qualidade das instituições de ensino, uma escola de boa qualidade pode acarretar uma boa formação cognitiva. Contudo, fatores não-cognitivos podem influenciar a proficiência dos alunos, mas não podem ser diretamente observados.

TABELA 7 - Resultado dos modelos econométricos estimados para função de produção educacional da nota da prova objetiva no ENEM 2008

	MQO	HECKMAN	HLM0	HLM1	HLM2
idade	-1,330***	-1,464***		-1,433***	-1,386***
mulher	-4,119***	-4,163***		-4,015***	-4,014***
nao_branco	-1,988***	-1,989***		-1,359***	-1,253***
trabalha	-1,579***	-1,586***		-1,063***	-0,987***
irmaos_m	-0,348***	-0,334***		-0,379***	-0,325***
noturno	-2,741***	-2,714***		-2,164***	-1,823***
cursinho	3,162***	3,144***		3,113***	3,078***
mae_naosei	-0,225	-0,272		-0,462	-0,554*
mae_fund1a4	0,560**	0,519*		0,509**	0,492**
mae_fund5a8	0,772***	0,702***		0,660***	0,614***
mae_em1	1,512***	1,440***		1,249***	1,140***
mae_em2	2,295***	2,201***		1,768***	1,572***
mae_sup1	4,383***	4,228***		3,172***	2,821***
mae_sup2	5,994***	5,839***		3,663***	3,296***
mae_pos	7,450***	7,282***		4,368***	4,018***
renda1sm	2,185***	1,828***		2,383***	2,431***
renda1a2sm	4,060***	3,848***		3,857***	3,742***
renda2a5sm	6,498***	6,189***		5,719***	5,273***
renda5a10sm	9,052***	8,695***		6,708***	5,919***
renda10sm	12,451***	11,982***		7,088***	6,488***
moradoresXrenda1sm	-0,483***	-0,457***		-0,420***	-0,395***
moradoresXrenda1a2sm	-0,510***	-0,524***		-0,413***	-0,385***
moradoresXrenda2a5sm	-0,428***	-0,418***		-0,307***	-0,248***
moradoresXrenda5a10sm	-0,380***	-0,368***		-0,146**	-0,08
moradoresXrenda10sm	-0,025	-0,01		0,165**	0,161*
esc_munic	1,799***	1,770***			0,831
esc_fed	18,839***	18,993***			21,754***
esc_priv	8,207***	8,481***			10,056***
reg_sa	-3,723***	-3,691***			-3,977***
reg_re	-4,944***	-5,095***			-5,446***
reg_rj	-4,008***	-4,304***			-4,037***
reg_sp	-3,064***	-3,188***			-2,908***
reg_pa	-1,522***	-1,449***			-1,045***
Razão Inversa de Mills		1,442***			
constante	44,929***	43,775***	44,166***	44,141***	43,700***
Mills (rho)		0,11			
ICC			40,59%	24,27%	13,78%
Nº Observações	159104	153134	173567	159104	159104

Nota: *: significante a 10%; **: significante a 5%; ***: significante a 1%.

Fonte dos dados básicos: INEP - ENEM 2008.

TABELA 8 - Resultado dos modelos econométricos estimados para função de produção educacional da nota da prova de redação do ENEM 2008

	MQO	HECKMAN	HLM0r	HLM1r	HLM2r
idade	-1,339***	-1,162***		-1,313***	-1,288***
mulher	2,907***	2,919***		2,933***	2,916***
nao_branco	-0,765***	-0,732***		-0,642***	-0,482***
trabalha	-0,709***	-0,547***		-0,367***	-0,387***
irmaos_m	-0,135	-0,11		-0,200**	-0,139
noturno	-1,577***	-1,353***		-1,661***	-1,455***
cursinho	0,820***	1,653***		1,491***	1,488***
mae_naosei	-0,475	-0,153		-0,329	-0,442
mae_fund1a4	1,181***	0,964***		1,221***	1,164***
mae_fund5a8	1,135***	0,837***		1,013***	0,964***
mae_em1	1,393***	1,060***		1,193***	1,113***
mae_em2	1,787***	1,424***		1,429***	1,306***
mae_sup1	2,039***	1,874***		1,815***	1,516***
mae_sup2	2,031***	2,229***		1,718***	1,369***
mae_pos	2,298***	2,828***		1,807***	1,477***
renda1sm	1,702***	1,831***		2,210***	2,139***
renda1a2sm	1,675***	2,465***		2,640***	2,274***
renda2a5sm	3,273***	3,603***		3,874***	3,259***
renda5a10sm	3,609***	3,864***		4,071***	3,034***
renda10sm	5,780***	5,048***		4,660***	3,662***
moradoresXrenda1sm	-0,200**	-0,404***		-0,364***	-0,295***
moradoresXrenda1a2sm	-0,120**	-0,261***		-0,168***	-0,124**
moradoresXrenda2a5sm	-0,258***	-0,247***		-0,211***	-0,161***
moradoresXrenda5a10sm	-0,274***	-0,176**		-0,194**	-0,121
moradoresXrenda10sm	-0,517***	-0,122		-0,215**	-0,209**
esc_munic	0,605**	0,628***			0,672
esc_fed	8,475***	8,458***			9,354***
esc_priv	2,167***	3,541***			3,408***
reg_sa	0,784***	0,973***			0,388
reg_re	-12,954***	-3,673***			-9,571***
reg_rj	-0,499***	-0,998***			-0,893***
reg_sp	-0,929***	-0,967***			-0,784***
reg_pa	1,300***	1,308***			1,718***
Razão Inversa de Mills		0,403***			
constante	57,251***	56,826***	58,913***	55,476***	55,842***
Mills (rho)		0,034			
ICC			13,13%	10,05%	6,91%
Nº Observações	159159	149145	173628	159159	159159

Nota: *: significativa a 10%; **: significativa a 5%; ***: significativa a 1%.

Fonte dos dados básicos: INEP - ENEM 2008.

6 A RELAÇÃO ENTRE PROFICIÊNCIA COGNITIVA E SITUAÇÃO OCUPACIONAL DOS JOVENS

Os objetivos deste capítulo são, em primeiro lugar, apresentar os resultados da imputação da proficiência cognitiva, medida pela nota do ENEM, para os indivíduos que estão terminando o ensino médio na base de dados da PME. Com isso, avaliamos a consistência do resultado mediante os dados utilizados.

Em segundo lugar, investigamos a situação ocupacional desses jovens doze meses após a primeira entrada na pesquisa da PME, explorando a proficiência estimada como fator associado. Com isso, analisamos se esta variável influencia a inserção dos jovens no mercado de trabalho, como proposto por Hanushek (2009) e analisado por Soares (2010).

6.1 Imputação da Proficiência Cognitiva do ENEM para a PME

A imputação da nota do Enem para os indivíduos na PME foi feita conforme a metodologia de Elbers et al. (2003) e utilizou as variáveis descritas anteriormente e que são comuns a ambas as bases de dados. Apesar da nota do ENEM estar no intervalo de 0 a 100 na escala do exame, nada garante que esses limites sejam respeitados quando estimamos a variável com outra amostra.

Por essa razão, os valores da nota imputada foram truncados para esses limites, mesmo porque tais casos foram minoria em nossa estimação, com apenas uma observação abaixo do limite inferior e nove acima do limite superior, de um total de 1838 observações. Os coeficientes estimados na base de dados do ENEM e que foram utilizados na imputação da nota na PME podem ser conferidos na tabela 9, onde foi estimado uma regressão para cada idade:

TABELA 9 - Resultado da estimação do modelo (3.17) para a imputação da nota do ENEM para os indivíduos da PME por idade

	Idade							
	17 anos	18 anos	19 anos	20 anos	21 anos	22 anos	23 anos	24 anos
mulher	-4,04 (0,101)	-3,769 (0,116)	-4,40 (0,226)	-3,628 (0,342)	-3,536 (0,469)	-3,049 (0,601)	-3,602 (0,721)	-3,871 (0,82)
nao_branco	-2,915 (0,109)	-2,621 (0,121)	-1,627 (0,235)	-1,05 (0,364)	-0,659 (0,508)	-1,497 (0,65)	0,028 (0,757)	-2,631 (0,866)
irmaos_m	-0,714 (0,132)	-0,546 (0,152)	0,210 (0,276)	-0,629 (0,416)	-0,51 (0,546)	1,322 (0,700)	-1,845 (0,835)	-2,011 (0,952)
so_mae	0,689 (0,121)	0,741 (0,134)	1,096 (0,240)	1,263 (0,361)	1,1 (0,507)	0,856 (0,660)	2,406 (0,820)	1,29 (0,921)
moradores	-1,206 (0,527)	-0,507 (0,494)	-1,105 (0,654)	-0,241 (0,864)	-0,706 (0,831)	-2,729 (1,408)	0,851 (1,137)	2,248 (2,558)
mae_fund1a4	1,533 (0,535)	0,546 (0,449)	-0,354 (0,614)	0,988 (0,766)	0,315 (0,952)	1,858 (1,062)	1,409 (1,127)	1,33 (1,237)
mae_fund5a8	2,114 (0,531)	0,865 (0,447)	-0,139 (0,619)	1,123 (0,781)	1,555 (0,978)	0,984 (1,135)	2,107 (1,206)	2,179 (1,347)
mae_em1	3,272 (0,544)	2,148 (0,469)	1,014 (0,678)	3,102 (0,881)	2,79 (1,136)	3,054 (1,391)	4,714 (1,602)	3,087 (1,889)
mae_em2	4,95 (0,526)	3,64 (0,445)	1,921 (0,622)	3,423 (0,797)	3,021 (1,042)	2,507 (1,217)	4,516 (1,412)	3,215 (1,677)
mae_sup1	8,178 (0,555)	6,983 (0,504)	5,170 (0,802)	6,877 (1,293)	6,281 (1,745)	3,066 (2,316)	5,824 (3,842)	4,309 (3,151)
mae_sup2	10,719 (0,538)	9,504 (0,472)	7,022 (0,728)	6,165 (1,087)	4,705 (1,652)	8,404 (2,235)	3,898 (2,654)	2,813 (3,637)
mae_pos	12,364 (0,563)	11,066 (0,516)	7,742 (0,889)	3,301 (1,815)	9,801 (2,740)	10,245 (4,368)	-4,243 (11,394)	0,202 (5,165)
renda1sm	-6,055 (2,187)	0,166 (2,023)	-0,373 (2,839)	3,350 (3,536)	-0,418 (3,361)	-2,714 (4,938)	5,366 (4,631)	5,032 (7,885)
renda1a2sm	-2,537 (2,136)	2,423 (1,978)	1,610 (2,795)	5,188 (3,489)	-0,569 (3,306)	-4,152 (4,862)	9,983 (4,542)	8,944 (7,804)
renda2a5sm	2,872 (2,132)	7,231 (1,979)	6,666 (2,824)	7,330 (3,564)	3,066 (3,464)	-2,888 (5,080)	7,54 (4,800)	10,695 (7,937)
renda5a10sm	8,497 (2,161)	14,272 (2,035)	9,705 (3,041)	13,259 (4,186)	5,077 (5,084)	-3,21 (6,968)	6,956 (8,647)	14,844 (10,098)
renda10sm	11,79 (2,172)	17,042 (2,056)	11,148 (3,164)	17,869 (5,032)	1,826 (8,697)	4,992 (12,438)	26,818 (18,722)	21,998 (12,900)
moradoresXrenda1sm	0,88 (0,545)	-0,118 (0,510)	0,594 (0,675)	-0,003 (0,890)	0,231 (0,872)	1,753 (1,447)	-0,963 (1,198)	-2,146 (2,600)
moradoresXrenda1a2sm	0,661 (0,533)	-0,139 (0,499)	0,591 (0,664)	-0,136 (0,878)	0,581 (0,855)	2,249 (1,425)	-1,629 (1,173)	-2,512 (2,572)
moradoresXrenda2a5sm	0,386 (0,532)	-0,283 (0,499)	0,112 (0,669)	-0,001 (0,890)	0,238 (0,881)	2,256 (1,455)	-0,66 (1,217)	-2,112 (2,590)
moradoresXrenda5a10sm	0,436 (0,538)	-0,635 (0,511)	0,262 (0,718)	-0,741 (1,018)	0,372 (1,204)	2,247 (1,818)	-0,329 (1,846)	-2,86 (2,873)
moradoresXrenda10sm	1,281 (0,541)	0,496 (0,516)	1,163 (0,748)	-1,340 (1,198)	-0,147 (1,996)	0,944 (2,704)	-3,036 (5,089)	-4,596 (3,298)
reg_sa	-1,935 (0,268)	-2,442 (0,279)	-2,123 (0,458)	-2,027 (0,673)	-3,014 (0,901)	-2,8 (1,195)	-3,335 (1,528)	-3,257 (1,601)
reg_re	-1,529 (0,233)	-3,874 (0,261)	-3,212 (0,462)	-2,825 (0,694)	-2,917 (0,931)	-3,671 (1,245)	-1,882 (1,463)	-0,737 (1,682)
reg_rj	-1,59 (0,191)	-2,974 (0,190)	-1,133 (0,355)	-0,336 (0,572)	-1,413 (0,810)	-1,328 (1,060)	-2,316 (1,297)	-1,731 (1,349)
reg_sp	-4,139 (0,162)	-4,764 (0,167)	-4,424 (0,357)	-2,649 (0,586)	-2,264 (0,803)	-1,907 (1,028)	-1,888 (1,197)	-3,354 (1,236)
reg_pa	-2,58 (0,237)	-3,676 (0,313)	0,409 (0,533)	2,507 (0,793)	2,174 (1,035)	3,511 (1,346)	2,394 (1,533)	2,713 (1,811)
_cons	50,29 (2,174)	45,382 (2,003)	41,473 (2,826)	32,944 (3,520)	37,361 (3,380)	41,474 (4,916)	29,479 (4,687)	31,263 (7,885)
N	77505	56594	12619	4549	2342	1435	1043	801

Fonte básica dos dados: ENEM 2008, PME 2008 e 2009

A tabela 10 mostra o resultado da imputação da nota do ENEM 2008 na PME 2008 e 2009 e que foi descrita pela variável “profic”. A média amostral do ENEM ficou entre as médias amostrais da nota imputada para os anos de 2008 e 2009 da PME, com um desvio-padrão um pouco maior.

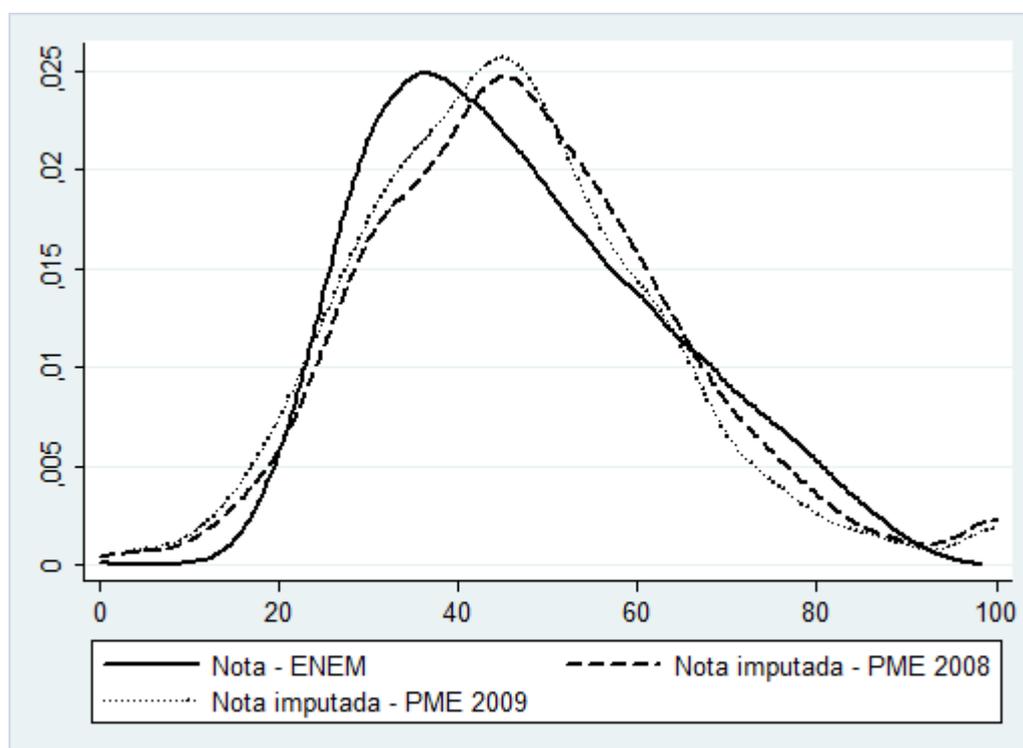
TABELA 10 – Estatísticas descritivas da nota do ENEM e da proficiência imputada na bases de dados da PME

Base de Dados	Variável	Observações	Média	Desvio Padrão	Mínimo	Máximo
ENEM-2008	nota	158469	46,80	16,52	0,00	98,41
PME-2008	profic	954	47,82	17,49	0,00	100,00
PME-2009	profic	884	45,90	16,95	2,36	100,00

Fonte: Elaboração própria com dados do ENEM 2008, PME 2008 e 2009

O gráfico 9 aponta para a existência de uma diferença na distribuição de proficiência entre as amostras. Pelo gráfico, podemos observar que o valor modal da nota imputada é maior que a calculada na base do ENEM. Além disso, notamos que existe uma concentração de indivíduos que obtiveram a sua proficiência estimada ao máximo, o que não é observado na base original.

GRÁFICO 9 - Distribuição da nota do ENEM 2008 e da nota imputada na PME 2008-2009



Fonte dos dados básicos: ENEM 2008, PME 2008 e 2009

Essa distorção pode ser resultado da distribuição desproporcional das variáveis dependentes em ambas as bases de dados, conforme foi discutido ao final da seção anterior. Aparentemente, os indivíduos menos proficientes estão em maior número na amostra do ENEM, mas, por exemplo, existem muito mais indivíduos com renda familiar elevada na PME do que na base de dados do ENEM e, como essa variável possui impacto positivo sobre a proficiência, a nota estimada para os indivíduos na PME tende a ser maior, considerando este efeito parcial. O mesmo acontece pela participação diferenciada de indivíduos nas regiões metropolitanas, que pode ser fruto do viés de auto-seleção da amostra do ENEM.

6.2 Análise dos Determinantes da Mobilidade Ocupacional dos Jovens

O principal objetivo deste capítulo é investigar se a proficiência cognitiva funciona como um determinante da mobilidade ocupacional dos jovens, no período de 12 meses, sendo que entre as possíveis situações ocupacionais, o jovem pode deixar de estudar e estar economicamente inativo (que inclui não trabalhar nem procurar emprego), pode estar no mercado de trabalho e deixar de estudar, pode estar ativo no mercado de trabalho e estudando ou, por fim, pode dedicar-se apenas aos estudos e ficar inativo no mercado de trabalho.

A estimação do modelo logit multinomial foi feita separadamente para homens e mulheres, considerando cada situação inicial como uma amostra à parte, o que resultou em quatro modelos estimados utilizando quatro subamostras¹⁵. Conforme mostrado pela descrição dos dados do quadro nº 4, existe uma tendência dos indivíduos transitarem para a situação de trabalho, deixando a escola, após o período de 12 meses.

Queremos investigar se a inclusão da variável indicativa da proficiência cognitiva altera a probabilidade de transição entre as categorias. Para controlar pelo ano

¹⁵ Em busca de um melhor ajuste dos dados, tentamos estimar o modelo de transição ocupacional utilizando outros recortes amostrais, separando os indivíduos por sexo e região metropolitana ou agregando pelas regiões sul, sudeste e nordeste. Apesar disso, os resultados foram insatisfatórios ou pouco informativos diante nossos de objetivos, com poucas observações em algumas categorias, o que resultava na não convergência de alguns modelos.

em que o indivíduo entrar na amostra, acrescentamos a variável binária “v2009”, que indica se o mesmo entrou na amostra em 2009 ou, caso contrário, em 2008. As tabelas 11 e 12 apresentam as razões de risco relativo estimadas pelos modelos econométricos, para as duas situações iniciais possíveis (ativo do mercado de trabalho e estuda; inativo no mercado de trabalho e estuda), respectivamente. Cada coluna representa o resultado de um modelo estimado para homens e mulheres.

A princípio, como observado na parte inferior das tabelas de resultados, o ajuste do modelo cuja subamostra foi composta por indivíduos os quais a situação inicial era estudar e trabalhar, mostrou-se pouco significativo. O p-valor do teste Qui-Quadrado, que compara o modelo estimado ao modelo nulo, sem variáveis, ficou bastante elevado, nessas amostras, tanto para homens (0,484), quanto para mulheres (0,715), quando comparados ao nível de 5% de significância.

Em outras palavras, por exemplo, ao nível de 5% de significância, não é possível rejeitar a hipótese nula, de que os coeficientes do modelo são todos simultaneamente iguais a zero, não havendo efeito estatisticamente significativo pelas variáveis independentes. Ou seja, a proficiência cognitiva, nesse caso, não possui efeito algum sobre o risco relativo de mudar de situação, ela não é um agente de mudança para os indivíduos, para nenhuma situação.

Com respeito aos modelos estimados para a subamostra daqueles indivíduos que apenas estudavam no período inicial (não trabalha e estuda), o teste Qui-Quadrado rejeita a hipótese nula, tanto para homens quanto para mulheres. Assim, podemos afirmar que pelo menos um dos coeficientes de cada modelos difere estatisticamente de zero, ambos ao nível de 5% de significância.

A razão de risco relativo mostra se, para um aumento marginal na proficiência cognitiva, o indivíduo está mais (ou menos) propenso a transitar da categoria de referência. Quando o seu valor do coeficiente da variável é maior do que a unidade, então existe uma propensão de mobilidade da categoria de referência para a outra situação, considerando a variação de um ponto na escala de proficiência estimada pela nota do ENEM. Se o coeficiente for menor do que a

unidade, a propensão de que o indivíduo permaneça na categoria de referência é maior, quanto maior a proficiência.

TABELA 11 - Razão de risco relativo para as situações ocupacionais estimadas pelo modelo logit multinomial para a situação inicial "Trabalha e Estuda"

Variável dependente (situação em t+12)	Variáveis independentes	Sexo	
		Homens	Mulheres
Não trabalha e não estuda	profic	0,987 (0,012)	0,996 (0,011)
	v2009	0,538 (0,216)	1,072 (0,389)
	constante	1,133 (0,696)	1,200 (0,668)
Trabalha e não estuda	profic	0,992 (0,008)	0,989 (0,009)
	v2009	0,686 (0,181)	1,010 (0,301)
	constante	4,360*** (1,833)	4,979*** (2,272)
Trabalha e estuda	profic		
	v2009	<i>categoria de referência</i>	<i>categoria de referência</i>
	constante		
Não trabalha e estuda	profic	1,009 (0,015)	1,006 (0,012)
	v2009	0,694 (0,349)	1,100 (0,476)
	constante	0,193*** (0,156)	0,399 (0,262)
N		347	354
Graus de liberdade		6	6
Valor χ^2		5,48	3,72
P-valor do teste χ^2		0,484	0,715

Legenda: * p < 10%, ** p < 5%, *** p < 1%

Fonte básica dos dados: ENEM 2008, PME 2008 e 2009

TABELA 12 - Razão de risco relativo para as situações ocupacionais estimadas pelo modelo logit multinomial para a situação inicial "Não Trabalha e Estuda"

Variável dependente (situação em t+12)	Variáveis independentes	Sexo	
		Homens	Mulheres
Não trabalha e não estuda	profic	0,987** (0,006)	0,986** (0,006)
	v2009	0,927 (0,218)	1,432* (0,306)
	constante	1,726 (0,654)	1,412 (0,475)
Trabalha e não estuda	profic	0,980*** (0,006)	0,981*** (0,006)
	v2009	1,046 (0,233)	1,237 (0,256)
	constante	2,905*** (1,047)	2,174** (0,706)
Trabalha e estuda	profic	0,992 (0,009)	1,004 (0,008)
	v2009	0,900 (0,291)	1,024 (0,300)
	constante	0,508 (0,264)	0,265*** (0,122)
Não trabalha e estuda	profic		
	v2009	<i>categoria de referência</i>	<i>categoria de referência</i>
	constante		
N		525	612
Graus de liberdade		6	6
Valor χ^2		12,91	16,31
P-valor do teste χ^2		0,044	0,012

Legenda: * p < 10%, ** p < 5%, *** p < 1%

Fonte básica dos dados: ENEM 2008, PME 2008 e 2009

Ao considerar apenas os modelos estimados para os indivíduos que apenas estudavam, aqueles que se apresentaram estatisticamente significativos no teste qui-quadrado, nas tabelas 11 e 12, podemos observar que, quanto maior a proficiência, menor o risco relativo do indivíduo passar para uma situação em que ele não estuda. A afirmação anterior fica comprovada pelos valores dos coeficientes da proficiência menores que a unidade para a primeira e segunda categoria da variável dependente, para ambos os sexos.

A razão de risco relativo devida ao aumento de um ponto na escala de proficiência cognitiva estimada para a situação (não trabalha e não estuda) relativa à situação (não trabalha e estuda), dado que as outras variáveis foram mantidas constantes, foi de 0,987 para os homens e de 0,986 para as mulheres.

Em outras palavras, com aumento da proficiência cognitiva, aumenta-se a expectativa de que o indivíduo permaneça na situação de referência, o que parece bastante razoável, uma vez que ele estará ocupando todo o seu tempo à continuação de sua formação escolar.

A interpretação do coeficiente da proficiência cognitiva para a situação “trabalha e não estuda”, tendo como referência a situação “não trabalha e estuda” é semelhante. Como o valor é inferior a um, para ambos os sexos, significa que quanto maior a proficiência cognitiva, menores são as chances de que o jovem deixe de estudar para apenas trabalhar.

O coeficiente da variável proficiência para a situação do jovem trabalhar e estudar não apresentou um valor estatisticamente diferente de um. Isto é, um aumento na proficiência cognitiva não aumenta ou diminui o risco relativo de um indivíduo que estava cursando o último ano do ensino médio e não trabalhava, passar a trabalhar e continuar a sua educação formal.

Assim, o valor dos coeficientes da proficiência cognitiva no modelo logit multinomial estimado para explicar a mobilidade ocupacional dos jovens, pelo menos no prazo de doze meses, pode indicar como verdadeira a hipótese de que o aumento da proficiência cognitiva pode induzir a situação ocupacional dos jovens.

A matriz de transição apresentada no quadro N^o4 mostrou que grande parte dos jovens cessa os estudos após a conclusão do ensino médio. No entanto, o modelo estimado propõe que, quanto maior a proficiência cognitiva, maior a probabilidade de o jovem continuar estudando e, conseqüentemente, acumulando capital humano.

As bases de dados utilizadas são consistentes, no sentido de que servem à finalidade a que se destinam, apesar do ENEM apresentar um viés de auto-seleção, não representando fielmente a população composta por todos os concluintes do ensino médio, o que pode afetar os nossos resultados e distorcer, pelo menos em parte, a nossa análise. A metodologia de compatibilização das bases buscou apresentar uma alternativa eficaz para estudar como a proficiência cognitiva induz a situação ocupacional dos jovens.

Os resultados diferem daqueles obtidos por Soares (2010, p.106-110), que também utilizou o modelo logit multinomial para explicar a situação ocupacional dos jovens, utilizando a PME (2003-2008) e o ENEM (2003-2007), mas valeu-se de outra metodologia de compatibilização para imputação da nota do ENEM para os indivíduos da PME. O autor concluiu que um aumento na nota do ENEM aumenta as chances de estar empregado, após o término do ensino médio, mas não possui influência sobre a probabilidade do jovem entrar no ensino superior ou estar em um curso preparatório para vestibulares.

Os resultados poderiam ter uma significância estatística maior, ou até mesmo um efeito diferente, se houvesse a possibilidade de acompanhar os indivíduos por um período superior a doze meses em uma base longitudinal, considerando, assim, a situação ocupacional no longo prazo. Utilizamos a base de dados do ENEM 2008, pois ela foi a base mais atualizada com disponibilidade de dados relativos a pessoas que estão terminando o ensino médio. Contudo, uma possibilidade de extensão deste trabalho seria investigar o problema aqui levantado por um período maior, utilizando várias edições do ENEM e da PME.

7 CONCLUSÃO

A proficiência cognitiva é um fator importante na vida dos indivíduos e o seu nível é determinado por fatores pessoais, familiares e escolares, entre outros. Os trabalhos resenhados no segundo capítulo mostraram a relevância do tema para entender a desigualdade de renda entre as pessoas sob a luz da teoria do Capital Humano e a crítica da teoria da Sinalização sobre o papel da escola.

Por meio do método de mínimos quadrados ordinários, foi possível trabalhar em torno do nosso primeiro objetivo, o de estimar e analisar a função de produção educacional, utilizando a base de dados do ENEM 2008. Com o método de Heckman em dois estágios, testamos e não rejeitamos a hipótese do viés de seleção da amostra. Ademais, buscamos corrigi-lo. Por fim, os modelos hierárquicos estimados foram úteis para observar que as escolas possuem um efeito sobre todos os seus alunos, mas que a diferença da proficiência entre as instituições se deve, em grande parte, a características pessoais e familiares dos alunos.

A análise dos modelos estimados possibilitou verificar que a renda familiar foi o determinante mais importante da nota do ENEM, juntamente com a escolaridade da mãe. Dentre os fatores pessoais, destacamos que os negros/pardos possuem um desempenho inferior aos brancos/amarelos. As mulheres possuem uma desvantagem na prova objetiva, mas possuem vantagem na prova de redação, em relação ao sexo masculino.

A nota do ENEM, tomada como medida de proficiência cognitiva, foi imputada para os indivíduos na base de dados da PME, seguindo a metodologia de criação de mapas de pobreza. O resultado foi satisfatório, tendo em vista a carência de trabalhos com o mesmo objetivo, principalmente pela inexistência de uma base de dados que alie tanto informações sobre proficiência cognitiva quanto variáveis de *status* ocupacional.

Com respeito ao objetivo principal desta dissertação, a investigação do efeito da proficiência cognitiva do jovem sobre a sua situação ocupacional, o resultado obtido confirmou aquilo que já foi discutido pela teoria.

Os resultados mostram que quanto maior a proficiência cognitiva, maiores são as chances dos indivíduos continuarem os estudos no ensino superior, e isso é verdade para os jovens que não trabalhavam durante o término do ensino médio, isto é, maiores as chances de manter o *status* de estudante. Assim, os indivíduos com maior proficiência cognitiva tenham, talvez, um custo de oportunidade menor que os demais, em relação a uma atividade que envolva aprendizado.

Em face aos determinantes da proficiência cognitiva, vimos que a renda familiar possui grande influencia, de forma que alguns jovens com baixa renda familiar têm suas escolhas limitadas muito antes da época em que devem prestar o vestibular. Por esta razão, o Estado poderia intensificar o investimento na qualidade do ensino básico, como forma de competir com o ensino privado e garantir aos alunos da rede pública de ensino, que concentra a maioria dos jovens de baixa renda, chances de competir pelo seu futuro educacional.

Por outro lado, a escolha de interromper os estudos e entrar no mercado de trabalho pode ser aquela que os jovens realmente desejem. Neste contexto, para o jovem que faz esta opção, o Estado deveria reorientar o conteúdo programático do ensino básico, que, atualmente, é excessivamente científico-acadêmico, (exceto no caso de escolas técnicas), desenvolvendo habilidades voltadas diretamente para mercado de trabalho. Por exemplo, alguns países oferecem um conteúdo mais técnico, como é o caso da Alemanha, que permite melhor adaptação à dinâmica do mercado de trabalho. Sendo assim, o Estado poderia intensificar o investimento também em educação técnica, não só nos institutos federais e estaduais, mas promover disciplinas vocacionais, como atividades extracurriculares, para as redes públicas de ensino.

Por fim, cabe ressaltar que existem poucos trabalhos fundamentados na compatibilização de bases de dados para estudar a situação ocupacional dos jovens no Brasil com base na proficiência cognitiva. Os resultados sugerem que a

sua acumulação possui diversos determinantes, que podem refletir na sua situação ocupacional e nas oportunidades de gerações futuras.

A inexistência de uma base de dados que alie informações de proficiência cognitiva e mercado de trabalho foi uma das causas para as várias hipóteses que foram assumidas como verdadeiras neste trabalho. Como proposta de uma agenda futura de pesquisa, propomos avançar em estudos empíricos que aprimorem a compatibilização de bases de dados educacionais e de mercado de trabalho, bem como testes que possibilitem discutir a robustez de tais medidas.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBERNAZ, A.; FERREIRA, F.H.G.; FRANCO, C. Qualidade e equidade no ensino fundamental brasileiro. 2002. Disponível em: <http://www.ppe.ipea.gov.br/index.php/pppe/article/view/139/74>. Acesso em: 09/05/2011.

ANDRADE, R.J. **Qualidade e equidade na educação básica brasileira: as evidências do Saeb 1995-2003**. 2008. 230f. Dissertação (Mestrado em Educação) – Faculdade de Educação, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

ANDRADE, J.M.; LAROS, J.A. Fatores associados ao desempenho escolar: Estudo Multinível com os dados do SAEB/2001. 2007. Disponível em: < <http://www.scielo.br/pdf/ptp/v23n1/a05v23n1.pdf>>. Acesso em: 24 fev 2012.

ASSUNÇÃO, J.J.; ALVES, L.S. Restrições de crédito e decisões intra-familiares. **RBE**, v.61, n.2, p. 201-229, abr./jun. 2007.

ASTONE, N.M.; MCLANAHAN, S.S. Family structure, residential mobility, and school dropout: a research note. 1994. Disponível em:< <http://www.jstor.org/stable/2061791>>. Acesso em 17/05/2011.

BARBOSA, M.E.F.; FERNANDES, C. A escola brasileira faz diferença? Uma investigação dos efeitos da escola na proficiência em matemática em alunos da 4ª série. 2001. Disponível em: <<http://www.educacion.gob.es/cide/espanol/investigacion/rieme/documentos/files/barbosa/Barbosa.pdf>>. Acesso em: 10/05/2011.

BARROS, R.P.; MACHADO, A.F.; MENDONÇA, R.S.P. A desigualdade da pobreza: estratégias ocupacionais e diferenciais por gênero. 1997. Disponível em: < http://desafios2.ipea.gov.br/pub/td/1997/td_0453.pdf >. Acesso em 11 jan. 2012.

BECKER, G.S. **Investment in Human Capital: A Theoretical Analysis**. The Journal of Political Economy, Vol. 70, No. 5, Part 2: Investment in Human Beings(Oct., 1962), pp. 9-49. Disponível em: < <http://www.jstor.org/stable/1829103>>. Acesso em: 16 fev 2011.

BEN-PORATH, Y. **The Production of Human Capital and the Life Cycle of Earnings**. The Journal of Political Economy, Vol. 75, No. 4, Part 1 (Aug., 1967), pp. 352-365. Disponível em: < <http://www.jstor.org/stable/1828596>>. Acesso em: 16 fev. 2011.

BLAUG, M. The economic value of education: studies in the economics of education. Editora Aldershot: 1992.

BOWLES, S., GINTIS, H., OSBORNE, M. **The Determinants of Earnings: A Behavioral Approach.** Journal of Economic Literature, Vol. 39, No. 4 (Dec., 2001), pp. 1137-1176. Disponível em: < <http://www.jstor.org/stable/2698522>>. Acesso em: 12 mar. 2011.

BRASIL, 1996. Lei nº 9394, de 20 de dezembro de 1996.

BRASIL, 2005. Exame Nacional do Ensino Médio (ENEM): Fundamentação Teórico- Metodológica. Disponível em: < <http://historico.enem.inep.gov.br> >. Acesso em: 17 jan. 2012.

BRASIL, 2007. **Pesquisa mensal de emprego:** Série relatórios metodológicos. Disponível em: < www.ibge.gov.br/home/estatistica/indicadores/.../srmv23pme.pdf >. Acesso em: 09 nov. 2011.

CAMERON, A.C.; TRIVEDI, P.K. **Microeconometrics:** Methods and Applications. Cambridge: Cambridge University Press, 2005. Cap.15. Multinomial models. p.490-528.

CARVALHO, R.L. **Casa, comida e roupa lavada: fatores associados à saída do jovem brasileiro do domicílio de origem.** 2009. 95f. Dissertação. CEDEPLAR, UFMG, Belo Horizonte.

CÉSAR, C.C.; SOARES, J.F. Desigualdades acadêmicas induzidas pelo contexto escolar. 2001. Disponível em: < http://www.abep.nepo.unicamp.br/docs/rev_inf/vol18_n1e2_2001/vol18_n1e2_2001_6artigo_97_110.pdf>. Acesso em: 10/05/2011.

CIOL, M.A.; HOFFMAN, J.M.; DUDGEON, B.J.; SHUMWAY-COOK, A.YORKSTON, K.M.; CHAN, L. Understanding the use of weights in the analysis of data from multistage surveys. 2006. Disponível em: < <http://www.sciencedirect.com/science/article/pii/S0003999305012840>>. Acesso em: 28 fev 2012

COLEMAN, J.S.; CAMPBELL, E.Q.; HOBSON, C.J.; MCPARTLAN, J.; MOOD, A.M.; WEINFELD, F.D.; YORK, R.L. Equality of educational opportunity. 1966. Disponível em: < <http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/06389>>. Acesso em: 18 out. 2011.

COLLIER, W.; VALBUENA, J.; ZHU, Y. **What determines post-compulsory educational choices?** Evidence from the longitudinal survey of young people in England. 2011. Disponível em: < ideas.repec.org/p/ukc/ukcedp/1112.html>. Acesso em: 13 jan. 2012.

CLARK, K.B.; SUMMERS, L.H. **The dynamics of youth unemployment.** 1982. Disponível em: < www.nber.org/chapters/c7875.pdf >. Acesso em: 03 jan. 2012.

CRESPO, A; REIS, M.C. O efeito diploma no Brasil. 2006. Disponível em: < http://www.ipea.gov.br/agencia/images/stories/PDFs/mercadodetrabalho/08Nota2_Anna_Mauricio.pdf>. Acesso em: 12/05/2011.

CRESPO, A; REIS, M.C. Sheepskin effects and the relationship between earnings and education: analyzing their evolution over time in Brazil. 2009. Disponível em: <http://www.scielo.br/scielo.php?pid=S0034-1402009000300002&script=sci_arttext>. Acesso em: 11/05/2011.

CURI, A.Z.; MENEZES-FILHO, N.A. **A relação entre o desempenho escolar e os salários no Brasil**. 2006. Disponível em: <www.anpec.org.br/encontro2005/artigos/A05A158.pdf>. Acesso em: 17 jan. 2012.

CURI, A.Z.; MENEZES-FILHO, N.A. Os determinantes das transições ocupacionais no mercado de trabalho brasileiro. 2004. Disponível em: <www.anpec.org.br/encontro2004/artigos/A04A141.pdf>. Acesso em: 19 jan 2012.

ECKSTEIN, Z.; WOLPIN, K.I. Why youth drop out of high school: the impact of preferences, opportunities, and abilities. **Econometrica**, v. 67, n.6, p. 1295-1339, nov. 1999.

ELBERS, C.; LANJOUW, J.O.; LANJOUW P. **Micro-level estimation of poverty and inequality**. 2003. Disponível em: <are.berkeley.edu/~ligon/.../elbers-et al03.pdf>. Acesso em: 21 ago. 2011.

FERRÃO, M.E.; BELTRÃO, K.I.; SANTOS, D.P. **Políticas de não-repetência e a qualidade da educação**: evidências obtidas a partir da modelagem dos dados da 4ª série do Saeb-99. 2002. Disponível em: <www.fcc.org.br/pesquisa/publicacoes/eae/arquivos/1179/1179.pdf>. Acesso em: 04 jan. 2012.

FERREIRA, R.A. **Desigualdade de desempenho escolar dos alunos do ensino fundamental do estado de São Paulo**: uma análise de decomposição. 2008. 85f. Dissertação (mestrado em economia) – Universidade de São Paulo, Ribeirão Preto, 2008.

FERREIRA, S.G.; VELOSO F.A. Mobilidade intergeracional de educação no Brasil. **Pesquisa e Planejamento Econômico**. v.33, n.3, dez. 2003.

FLEISCHHAUER, K.J. **A Review of Human capital Theory**: Microeconomics. University of St. Garden, 2007. Disponível em: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=957993>. Acesso em: 12 jan. 2011.

FLORI, P.M.; MENEZES-FILHO, N.A. Mobilidade entre ocupações e efeitos salariais. 2008. Disponível em: <www.anpec.org.br/encontro2008/artigos/200807171058080-.pdf>. Acesso em: 19 jan. 2012.

FRANCO, C.; ORTIGÃO, I.; ALBERNAZ, A.; BONAMINO, A.; AGUIAR, G.; ALVES, F.; SÁTYRO N. **Qualidade e equidade em educação**: reconsiderando o significado de “fatores intra-escolares”. 2007. Disponível em: <http://www.scielo.br/scielo.php?pid=S0104-40362007000200007&script=sci_arttext&tlng=pt>. Acesso em: 03/05/2011.

GHEZ, J. Education, the price of time, and lyfe-cycle consumption. In JUSTER, F.T. (Org.) **Education, Income and Human Behavior**. Mcgraw-Hill Books Company: 1975. p.295-312.

GOLDIN, C. The Human Capital Century and American Leadership: virtues of the past. Disponível em: <<http://www.economics.harvard.edu/faculty/goldin/files/humancap.pdf>>. Acesso em: 02 abr 2012.

GOLGHER, A. Diálogos com o Ensino Médio 5: A Escola de Ensino Médio no Brasil Analisada a Partir de Dados do INEP. 2010a. Disponível em: <www.cedeplar.ufmg.br/pesquisas/td/TD%20395.pdf>. Acesso em: 12 jan. 2012.

GOLGHER, A. Diálogos com o Ensino Médio 6: A Escola de Ensino Médio no Brasil Analisada a Partir de Dados do INEP. 2010b. Disponível em: <www.cedeplar.ufmg.br/pesquisas/td/TD%20396.pdf>. Acesso em: 12 jan. 2012.

GONÇALVES, F.O.; FRANÇA, M.T.A. Transmissão intergeracional de desigualdade e qualidade educacional: avaliando o sistema educacional brasileiro a partir do SAEB 2003. Ensaio: avaliação de políticas públicas educacionais. v.16, n. 61, p. 639-662, out./dez. 2008.

GREENE W.H. **Econometric analysis**. New Jersey: Prentice Hall, 2003. Cap.22. Limited dependent variable and duration models. p.756-802.

GUIMARÃES, R.R.M., RIOS-NETO, E.L.G. **Desigualdades de oportunidades educacionais**: seletividade e progressão por série no Brasil, 1986 a 2008. Belo Horizonte: UFMG/CEDEPLAR, 2010. Disponível em: <http://ufmg.academia.edu/RaquelGuimaraes/Talks/27454/Desigualdade_de_oportunidades_educacionais_Seletividade_e_progressao_por_serie_no_Brasil_1986_a_2008>. Acesso em: 12 mar. 2011.

HALEY, W.J. **Human Capital**: The Choice between Investment and Income. The American Economic Review, Vol. 63, No. 5 (Dec., 1973), pp. 929-944. Disponível em: <<http://www.jstor.org/stable/1813915>>. Acesso em: 16 fev 2011.

HANUSHEK, E.A. Sample selectivity and the validity of international student achievement tests in economic research. 2010. Disponível em: <http://www.sciencedirect.com/science?_ob=MIimg&_imagekey=B6V84-51C9JYF-7-1&_cdi=5860&_user=686413&_pii=S0165176510003605&_origin=gateway&_coverDate=02%2F28%2F2011&_sk=998899997&_view=c&_wchp=dGLzVzb-zSkzk&_md5=a3363b06c1a0a28700e20cd4a68ec47f&_ie=/sdarticle.pdf>. Acesso em: 13/05/2011.

HANUSHEK, E.A. **The trade-off between child quantity and quality**. The Journal of Political Economy, Vol. 100, No. 1 (Feb., 1992), pp. 84-117. Disponível em: <<http://www.jstor.org/stable/2138807>>. Acesso em: 11 mar. 2011.

HANUSHEK, E.A. The economic value of education and cognitive skills. In: SYKES, D.; SCHNEIDER, B.; PLANK, D.N. (Eds.) Handbook of education policy research. Washington: AERA. 2009. p.39-56.

HARDIE, J.H. **Making it?** Paths to achieving and not achieving aspirations in the transition to adulthood. 2009. Disponível em: <<http://paa2009.princeton.edu/download.aspx?submissionId=91113>>. Acesso em: 09 mar 2012.

HECKMAN, J.J. Sample selection bias as a specification error. 1979. Disponível em: <<http://www.jstor.org/stable/1912352>>. Acesso em: 18 fev. 2011.

HECKMAN, J.J., RUBINSTEIN, Y. **The Importance of noncognitive Skills: Lessons from GED Testing Program.** The American Economic Review, Vol. 91, No. 2, Papers and Proceedings of the Hundred Thirteenth Annual Meeting of the American Economic Association (May, 2001), pp. 145-149. Disponível em: <<http://www.jstor.org/stable/2677749>>. Acesso em: 11 abr. 2011.

HECKMAN, J.J.; MASTEROV, D.V. **The productivity argument for investing in young children.** 2005. Disponível em: <http://jenni.uchicago.edu/human-inequality/papers/Heckman_final_all_wp_2007-03-22c_jsb.pdf>. Acesso em: 19 jan. 2012.

HOX, J. **Multilevel analysis: Techniques and applications.** New Jersey: Laurence Erlbaum Associates, 2002. Cap. 1. Introduction to multilevel analysis, p. 1-10.

LAM, D. MARTELETO, L. A Dinâmica da Escolaridade das Crianças Brasileiras durante a Transição Demográfica: Aumento no Tamanho da Coorte versus Diminuição no Tamanho da Família. 2002. Disponível em: <http://www.abep.nepo.unicamp.br/docs/anais/pdf/2002/Com_EDU_ST16_Lam_teto.pdf>. Acesso em: 16/11/2010.

LAYARD, R., PSACHAROULOPOUS, G. **The Screening Hypothesis and the Returns to Education.** The Journal of Political Economy, Vol. 82, No. 5 (Sep. - Oct., 1974), pp. 985-998. Disponível em: <<http://www.jstor.org/stable/1829179>>. Acesso em: 13 abr. 2011.

MACHADO, A.F.; MORO, S.; MARTINS, L.; RIOS, J. **Qualidade do Ensino em Matemática: Determinantes do Desempenho de Alunos em Escolas Públicas Estaduais Mineiras.** 2008. Disponível em: <http://www.anpec.org.br/revista/vol9/vol9n1p23_45.pdf>. Acesso em: 04 mai. 2011.

MARTELETO, L. **O papel do tamanho da família na escolaridade dos jovens.** 2002. Revista Brasileira de Estudos de População, v.19, n.2, jul./dez. Disponível em: <http://www.abep.nepo.unicamp.br/docs/rev_inf/vol19_n2_2002/vol19_n2_2002_10_artigo_p159a178.pdf> Acesso em: 11 abr. 2011.

MEDEIROS, M., OSÓRIO, R. **Mudanças nas Famílias Brasileiras: A Composição dos Arranjos Familiares entre 1978 e 1998.** Disponível em: <http://www.abep.nepo.unicamp.br/docs/rev_inf/vol17_n1e2_2000/vol17_n1e2_2000_4artigo_67_85.pdf>. Acesso em: 01 mai. 2011.

MENEZES-FILHO, N. **Os Determinantes do Desempenho Escolar do Brasil.** 2007. Disponível em: <http://www.cepe.ecn.br/seminarioiv/download/menezes_filho.pdf>. Acesso em: 07/05/2011.

MINCER, J. **Investment in human Capital and Personal Income Distribution.** The Journal of Political Economy, Vol. 66, No. 4 (Aug., 1958), pp. 281-302. Disponível em: <<http://www.jstor.org/stable/1827422>> . Acesso em: 18 fev. 2011.

ORDINE, P.; ROSE, G. **Inefficient self-selection into education and wage inequality.** 2011. Disponível em: <http://www.aiel.it/bacheca/SASSARI/papers/ordine_rose.pdf> . Acesso em: 04 mar 2012.

PATRINOS, H.A., PSACHAROPOULOS, G. **Family Size, Schooling and Child Labor in Peru – An Empirical Analysis.** Disponível em: <<http://link.springer.de/link/service/journals/00148/papers/7010004/70100387.pdf>> . Acesso em: 11 abr. 2011.

PISA. **The Programme for International Student Assessment.** 2006. Disponível em: <<http://www.pisa.oecd.org>> . Acesso em 11 jan. 2012.

RIANI, J.L.R.; RIOS-NETO, E.L.G. Background familiar versus perfil escolar do município: qual possui maior impacto na resultado educacional dos alunos brasileiros?. 2008. Disponível em: <http://www.scielo.br/scielo.php?pid=S0102-30982008000200004&script=sci_arttext&tlng=es>. Acesso em: 17/05/2011.

RIBAS, R.P.; MACHADO, A.F. **A imputação da renda do não trabalho na pesquisa mensal de emprego.** 2009. Disponível em: <http://www.ppe.ipea.gov.br/index.php/ppe/article/viewFile/1157/1062>. Acesso em: 10 fev. 2011.

RIBAS, R.P.; SOARES, S.S.D. **Sobre o painel da pesquisa mensal de emprego (PME) do IBGE.** 2008. Disponível em: <www.ipea.gov.br/sites/000/2/publicacoes/tds/td_1348.pdf> . Acesso em: 08 ago. 2011.

SAMPAIO, C.E.M. Situação ocupacional dos jovens brasileiros na faixa etária de 15 a 17 anos. 2009. Disponível em: <<http://www.publicacoes.inep.gov.br>>. Acesso em: 12 mar 2012.

SCHOON I.; POLEK, E. Tenuage career aspirations and adult career attainment: the role of gender, social background and general cognitive ability. 2011. Disponível em: <<http://jbd.sagepub.com/content/35/3/210>>. Acesso em: 09 dez. 2011.

SCHULTZ, T.W. **Investment in Human Capital.** The American Economic Review, Vol. 51, No. 1 (Mar., 1961), pp. 1-17. Disponível em: <<http://www.jstor.org/stable/1818907>>. Acesso em: 16 fev 2011.

SILVA, N.V., HASENBALG, C. **Recursos Familiares e transições Educacionais**. Disponível em: < <http://www.scielo.br/pdf/csp/v18s0/13794.pdf>>. Acesso em: 01 mai. 2011.

SOARES, J.F. Avaliação da qualidade da educação escolar brasileira. 2008. Manuscrito. apud ANDRADE, R.J. **Qualidade e equidade na educação básica brasileira: as evidências do Saeb 1995-2003**. 2008. 230f. Dissertação (Mestrado em Educação) – Faculdade de Educação, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

SOARES, J.F. O efeito da escola sobre o desempenho cognitivo dos alunos. 2004. Disponível em: <<http://www.ppe.ipea.gov.br/index.php/ppe/article/view/139/74>>. Acesso em: 02/04/2011.

SOARES, S.S.D. **O bônus demográfico relativo e absoluto no acesso à escola**. 2008. Disponível em: <www.ipea.gov.br/sites/000/2/publicacoes/tds/td_1340.pdf>. Acesso em: 11 fev. 2011.

SOARES, S.S.D. **O Conteúdo Econômico da Educação**. 2010. 146f. Tese (Doutorado em Economia) – Programa de Pós-Graduação em Economia, Universidade de Brasília, Distrito Federal, 2010.

SOARES, T.M. **Modelo de três níveis hierárquicos para proficiência dos alunos de 4a série avaliados no teste de língua portuguesa do SIMAVE/PROEB-2002**. 2005. Disponível em: <<http://www.fcc.org.br/pesquisa/publicacoes/eae/arquivos/1112/1112.pdf>>. Acesso em: 08/04/2011.

SPENCE, M. **Job market Signalling**. The Quarterly Journal of Economics, Vol. 87, No. 3 (Aug., 1973), pp. 355-374. Disponível em: <<http://www.jstor.org/stable/1882010>>. Acesso em: 11 mar. 2011.

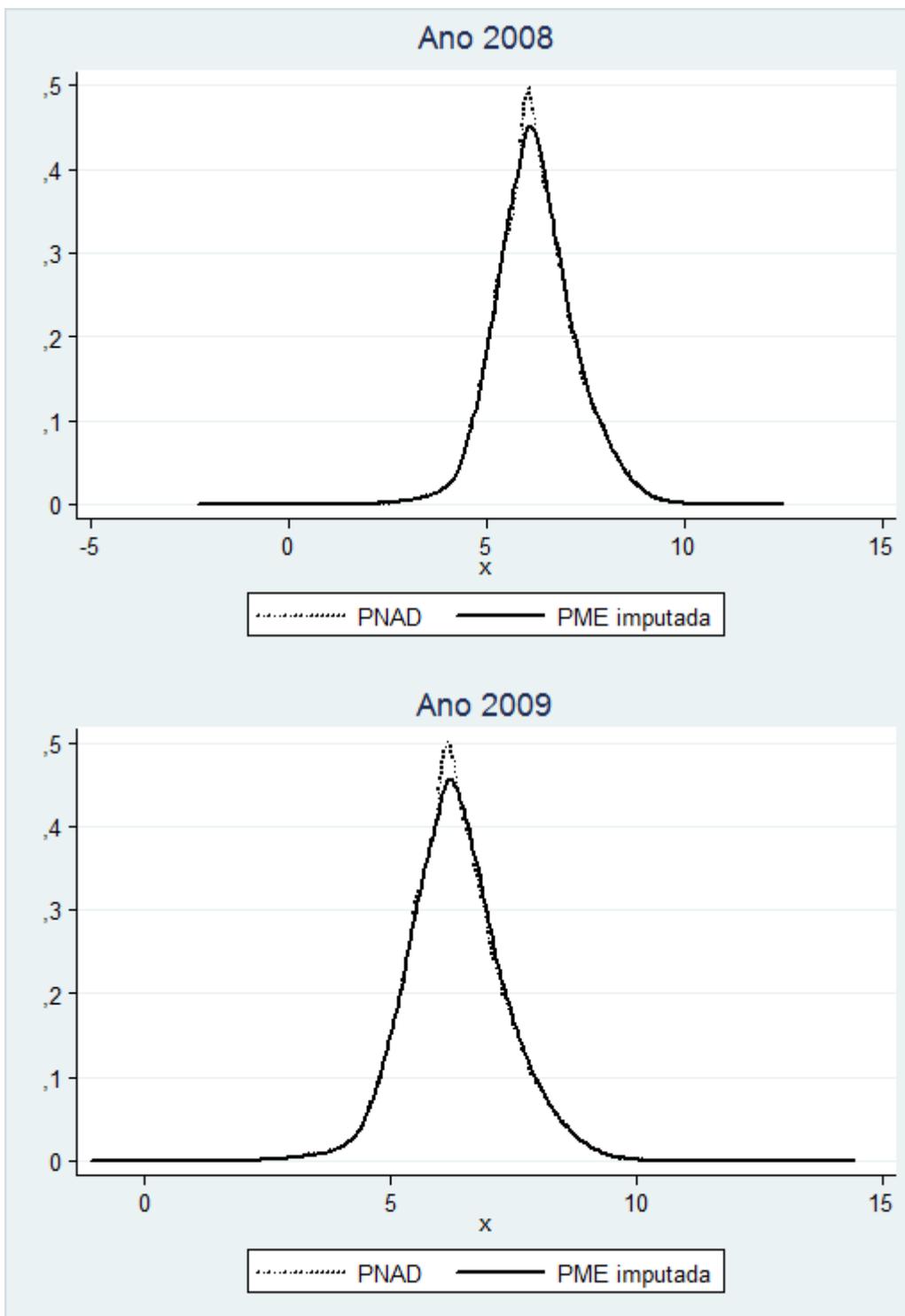
WOOLDRIDGE, J.M. **Econometric Analysis of cross section and panel data**. Cambridge: The MIT Press, 2001. Cap.4. The single-equation linear model and OLS estimation. p.49-76.

ANEXOS

Quadro A.1: Variáveis utilizadas no processo de imputação da nota do ENEM na PME

<i>Dummy</i> de sexo
<i>Dummy</i> de cor/raça
<i>Dummy</i> se mora com irmão
<i>Dummy</i> se mora apenas com a mãe
<i>Dummy</i> se a escolaridade máxima da mãe foi da 1ª a 4ª série do ensino fundamental
<i>Dummy</i> se a escolaridade máxima da mãe foi da 5ª a 8ª série do ensino fundamental
<i>Dummy</i> se a escolaridade máxima da mãe foi o ensino médio incompleto
<i>Dummy</i> se a escolaridade máxima da mãe foi o ensino médio completo
<i>Dummy</i> se a escolaridade máxima da mãe foi o ensino superior incompleto
<i>Dummy</i> se a escolaridade máxima da mãe foi o ensino superior completo
<i>Dummy</i> se a escolaridade máxima da mãe foi a pós-graduação
<i>Dummy</i> para renda familiar até 1 salário mínimo
<i>Dummy</i> para renda familiar de 1 a 2 s.m.
<i>Dummy</i> para renda familiar de 2 a 5 s.m.
<i>Dummy</i> para renda familiar de 5 a 10 s.m.
<i>Dummy</i> para renda familiar acima de 10 s.m.
Variável do número de moradores no domicílio
Variáveis de interação entre número de moradores e faixa de renda familiar
<i>Dummy</i> para a região metropolitana de salvador
<i>Dummy</i> para a região metropolitana de recife
<i>Dummy</i> para a região metropolitana de belo horizonte
<i>Dummy</i> para a região metropolitana de são paulo
<i>Dummy</i> para a região metropolitana de rio de janeiro
<i>Dummy</i> para a região metropolitana de porto alegre

GRÁFICO A.1 - Distribuição do log da renda per capita na base de dados da PNAD e PME



Fonte dos dados básicos: IBGE - PNAD 2008-2009 e PME 2008-2009

GRÁFICO A.2 - Distribuição percentual dos indivíduos segundo a região metropolitana em que reside, pela base de dados do ENEM e da PME com expansão da amostra

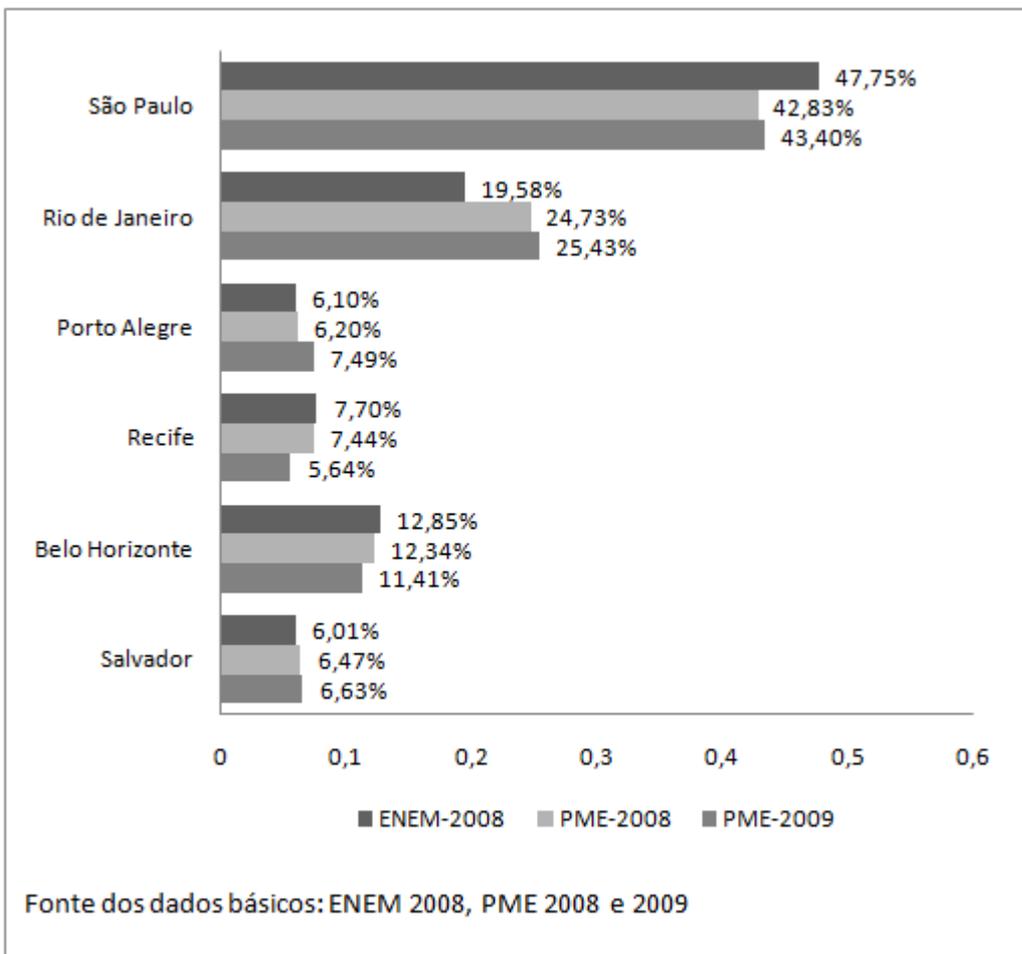


TABELA A.1 - Primeiro estágio de Heckman - modelo Probit para a probabilidade do indivíduo comparecer à prova objetiva do ENEM

Idade média dos docentes	-0,063*** (0,004)
Proporção de docentes com pós-graduação	0,074 (0,057)
Dummy para laboratório de informática	0,122** (0,052)
Dummy para laboratório de ciências	0,602*** (0,034)
Dummy para biblioteca	0,21*** (0,057)
Dummy para acesso à internet	0,765*** (0,109)
Dummy para banda larga na escola	0,148*** (0,039)
constante	0,518*** (0,181)
<hr/>	
N	7707
Graus de liberdade	7
Valor χ^2	934,58
P-valor do teste χ^2	0,000

Legenda: * p < 10%, ** p < 5%, *** p < 1%

Fonte básica dos dados: ENEM 2008, PME 2008 e 2009

TABELA A.2 - Primeiro estágio de Heckman - modelo Probit para a probabilidade do indivíduo comparecer à prova de redação do ENEM

Idade média dos docentes	-0,063*** (0,004)
Proporção de docentes com pós-graduação	0,072 (0,057)
Dummy para laboratório de informática	0,119** (0,052)
Dummy para laboratório de ciências	0,601*** (0,034)
Dummy para biblioteca	0,207*** (0,057)
Dummy para acesso à internet	0,765*** (0,109)
Dummy para banda larga na escola	0,148*** (0,039)
constante	0,518*** (0,181)
N	7696
Graus de liberdade	7
Valor χ^2	928,75
P-valor do teste χ^2	0,000

Legenda: * p < 10%, ** p < 5%, *** p < 1%

Fonte básica dos dados: ENEM 2008, PME 2008 e 2009

TABELA A.3 - Resultado do modelo (3.19) que estima o erro do modelo de imputação da nota do ENEM na base da PME por idade

	Idade							
	17 anos	18 anos	19 anos	20 anos	21 anos	22 anos	23 anos	24 anos
mulher	-0,191 (0,017)	-0,249 (0,020)	-0,369 (0,042)	-0,398 (0,070)	-0,550 (0,102)	-0,290 (0,118)	-0,655 (0,157)	-0,454 (0,165)
nao_branco	-0,105 (0,018)	-0,122 (0,020)	-0,111 (0,043)	-0,026 (0,074)	-0,201 (0,110)	-0,381 (0,128)	-0,034 (0,165)	-0,068 (0,175)
irmaos_m	-0,05 (0,022)	-0,026 (0,026)	-0,016 (0,051)	-0,114 (0,085)	-0,005 (0,119)	0,147 (0,138)	-0,093 (0,182)	-0,188 (0,192)
so_mae	0,021 (0,020)	-0,005 (0,023)	0,079 (0,044)	0,025 (0,074)	0,182 (0,110)	-0,165 (0,130)	-0,025 (0,178)	-0,046 (0,186)
moradores	-0,125 (0,088)	-0,195 (0,083)	-0,200 (0,121)	0,207 (0,176)	-0,191 (0,181)	0,042 (0,277)	0,418 (0,248)	0,377 (0,515)
mae_fund1a4	0,227 (0,089)	0,030 (0,076)	-0,261 (0,113)	-0,048 (0,156)	-0,283 (0,207)	0,343 (0,209)	0,411 (0,245)	0,109 (0,249)
mae_fund5a8	0,321 (0,089)	0,084 (0,075)	-0,268 (0,114)	-0,009 (0,159)	0,044 (0,213)	0,130 (0,223)	0,606 (0,262)	0,285 (0,271)
mae_em1	0,349 (0,091)	0,093 (0,079)	-0,127 (0,125)	0,332 (0,180)	0,213 (0,247)	0,104 (0,274)	0,622 (0,349)	0,732 (0,380)
mae_em2	0,462 (0,088)	0,263 (0,075)	0,051 (0,115)	0,492 (0,163)	0,509 (0,227)	0,505 (0,239)	1,009 (0,307)	0,39 (0,338)
mae_sup1	0,558 (0,093)	0,397 (0,085)	0,252 (0,148)	0,134 (0,264)	0,534 (0,380)	0,175 (0,456)	0,567 (0,836)	-0,426 (0,635)
mae_sup2	0,529 (0,090)	0,363 (0,080)	0,263 (0,134)	0,634 (0,222)	0,723 (0,359)	1,628 (0,440)	1,245 (0,578)	1,365 (0,733)
mae_pos	0,441 (0,094)	0,265 (0,087)	0,403 (0,164)	0,897 (0,371)	0,467 (0,596)	2,248 (0,859)	-61,21 (2,480)	2,282 (1,040)
renda1sm	-0,747 (0,366)	-0,818 (0,341)	-0,591 (0,523)	1,478 (0,722)	-0,583 (0,731)	-0,300 (0,971)	1,889 (1,008)	0,841 (1,589)
renda1a2sm	-0,627 (0,357)	-0,723 (0,333)	-0,485 (0,515)	1,308 (0,712)	-0,361 (0,719)	0,267 (0,956)	1,151 (0,989)	1,661 (1,572)
renda2a5sm	-0,471 (0,356)	-0,544 (0,334)	-0,150 (0,520)	1,648 (0,727)	0,239 (0,753)	-0,220 (0,999)	1,201 (1,045)	1,952 (1,599)
renda5a10sm	-0,46 (0,361)	-0,486 (0,343)	-0,024 (0,56)	1,835 (0,854)	-0,006 (1,106)	1,086 (1,371)	-0,923 (1,882)	3,804 (2,034)
renda10sm	-0,948 (0,363)	-0,55 (0,347)	-0,713 (0,583)	1,932 (1,027)	2,324 (1,892)	3,013 (2,446)	2,125 (4,075)	5,242 (2,599)
moradoresXrenda1sm	0,114 (0,091)	0,141 (0,086)	0,157 (0,124)	-0,311 (0,182)	0,222 (0,190)	0,044 (0,285)	-0,595 (0,261)	-0,337 (0,524)
moradoresXrenda1a2sm	0,091 (0,089)	0,149 (0,084)	0,191 (0,122)	-0,236 (0,179)	0,198 (0,186)	-0,029 (0,280)	-0,373 (0,255)	-0,529 (0,518)
moradoresXrenda2a5sm	0,106 (0,089)	0,166 (0,084)	0,141 (0,123)	-0,250 (0,182)	0,125 (0,192)	0,060 (0,286)	-0,305 (0,265)	-0,532 (0,522)
moradoresXrenda5a10sm	0,117 (0,090)	0,185 (0,086)	0,191 (0,132)	-0,260 (0,208)	0,186 (0,262)	-0,085 (0,358)	0,087 (0,402)	-0,774 (0,579)
moradoresXrenda10sm	0,186 (0,090)	0,164 (0,087)	0,352 (0,138)	-0,211 (0,245)	-0,279 (0,434)	-0,478 (0,532)	-0,29 (1,108)	-1,1 (0,664)
reg_sa	-0,099 (0,045)	-0,16 (0,047)	-0,207 (0,084)	-0,167 (0,137)	-0,323 (0,196)	-0,261 (0,235)	-0,073 (0,333)	0,243 (0,323)
reg_re	-0,054 (0,039)	-0,092 (0,044)	-0,23 (0,085)	-0,144 (0,142)	-0,317 (0,202)	-0,252 (0,245)	0,154 (0,318)	-0,016 (0,339)
reg_rj	-0,146 (0,032)	-0,181 (0,032)	-0,154 (0,065)	-0,114 (0,117)	-0,126 (0,176)	-0,422 (0,209)	0,453 (0,282)	-0,101 (0,272)
reg_sp	-0,13 (0,027)	-0,208 (0,028)	-0,248 (0,066)	-0,220 (0,120)	-0,187 (0,175)	-0,321 (0,202)	0,254 (0,260)	-0,003 (0,249)
reg_pa	-0,285 (0,040)	-0,264 (0,053)	-0,045 (0,098)	0,009 (0,162)	-0,378 (0,225)	-0,050 (0,265)	0,463 (0,334)	0,639 (0,365)
_cons	-3,841 (0,363)	-3,56 (0,338)	-3,204 (0,521)	-5,257 (0,718)	-3,753 (0,735)	-4,174 (0,967)	-5,831 (1,020)	-5,059 (1,588)
N	77505	56594	12619	4549	2342	1435	1043	801

Fonte básica dos dados: ENEM 2008, PME 2008 e 2009