https://doi.org/10.18222/eae.v36.11449

TRATAMENTO DE DADOS AUSENTES EM UMA AVALIAÇÃO EDUCACIONAL COM DADOS LONGITUDINAIS

LUIS GUSTAVO DO AMARAL VINHA[I]
JACOB ARIE LAROS[II]

[I] Universidade de Brasília (UnB), Brasília-DF, Brasil; *luisvinha@unb.br*
[II] Universidade de Brasília (UnB), Brasília-DF, Brasil; *jalaros@gmail.com*

RESUMO

A ausência de dados nas avaliações educacionais está relacionada com desempenho e perfil dos estudantes. O presente estudo propõe uma nova abordagem baseada em modelos de misturas de padrões para análise de dados incompletos em avaliações longitudinais. Essa abordagem é comparada com os procedimentos *listwise deletion* (LD) e imputação múltipla (IM), utilizando modelos de crescimento linear, com base em uma amostra de 8.681 estudantes do ensino médio do Ceará. Verifica-se que os procedimentos diferem na estimação dos efeitos das variáveis preditoras e da taxa média de aprendizado em matemática. Com o uso da nova abordagem são obtidas estimativas mais realistas para a taxa média de aprendizado e as trajetórias geradas são mais coerentes do que aquelas estimadas pelo procedimento de imputação múltipla.

PALAVRAS-CHAVE DESEMPENHO ACADÊMICO • TRATAMENTO DE DADOS AUSENTES • ESTUDO LONGITUDINAL • ANÁLISE DE REGRESSÃO.

COMO CITAR:

Vinha, L. G. do A., & Laros, J. A. (2025). Tratamento de dados ausentes em uma avaliação educacional com dados longitudinais. *Estudos em Avaliação Educacional*, *36*, Artigo e11449. https://doi.org/10.18222/eae.v36.11449

TRATAMIENTO DE DATOS AUSENTES EN UNA EVALUACIÓN EDUCATIVA CON DATOS LONGITUDINALES

RESUMEN

La ausencia de datos en las evaluaciones educativas está relacionada con el rendimiento y el perfil de los estudiantes. Este estudio propone un nuevo enfoque basado en modelos de patrones mezclados para analizar datos incompletos en evaluaciones longitudinales. Este enfoque es comparado con los procedimientos *listwise deletion* (LD) e imputación múltiple (IM), utilizando modelos de crecimiento lineal, a partir de una muestra de 8.681 estudiantes de educación secundaria de Ceará, Brasil. Se constata que los procedimientos difieren en la estimación de los efectos de las variables predictoras y de la tasa media de aprendizaje en matemáticas. Con el uso del nuevo enfoque, son obtenidas estimaciones más realistas para la tasa media de aprendizaje y las trayectorias generadas son más coherentes que aquellas estimadas por el procedimiento de imputación múltiple.

PALABRAS CLAVE RENDIMIENTO ACADÉMICO • TRATAMIENTO DE DATOS AUSENTES • ESTUDIO LONGITUDINAL • ANÁLISIS DE REGRESIÓN.

TREATMENT OF MISSING DATA IN AN EDUCATIONAL EVALUATION WITH LONGITUDINAL DATA

ABSTRACT

The absence of data in educational assessments is related to student's performance and profile. This study proposes a new approach based on pattern-mixture models for analyzing incomplete data in longitudinal evaluations. This approach is compared with listwise deletion (LD) and multiple imputation (IM) procedures, using linear growth models, based on a sample of 8,681 high school students in Ceará state, Brazil. The results show that the procedures differ in estimating the effects of predictor variables and average rate of learning in mathematics. The new approach yields more realistic estimates are obtained for the average rate of learning and the trajectories generated are more coherent than those estimated by the multiple imputation procedure.

KEYWORDS ACADEMIC PERFORMANCE • MISSING DATA TREATMENT • LONGITUDINAL STUDY • REGRESSION ANALYSIS.

Recebido em: 24 SETEMBRO 2024
Aprovado para publicação em: 5 AGOSTO 2025

INTRODUÇÃO

A ocorrência de dados ausentes é comum na pesquisa social aplicada (Occhipinti, 2024; Rousseau et al., 2012). A perda de informação pode ocorrer de diversas formas, e, em geral, o problema é maior nos estudos longitudinais, uma vez que os participantes podem se ausentar em qualquer momento da coleta dos dados (Enders, 2022; Little, 2024; McKnight et al., 2007). Levando em consideração os diferentes padrões de ausência e a disponibilidade de inúmeros métodos para a análise de dados incompletos, a escolha da abordagem mais adequada para cada situação não é uma tarefa fácil.

Existem muitas razões para a perda de informação nas pesquisas de levantamento (Enders, 2022; McKnight et al., 2007; Occhipinti, 2024). Os valores ausentes podem ser ocasionados por falhas no planejamento ou na execução da pesquisa, por exemplo: parte dos indivíduos selecionados não tem o conhecimento necessário para responder a alguns itens do questionário; itens que abordam temas delicados podem constranger o participante, gerando não respostas; questionários longos e cansativos podem ser abandonados antes do final; ou, ainda, falhas na transcrição dos dados para planilhas podem acontecer. Fatores relacionados aos participantes também podem gerar perda de informação, como nos casos em que um indivíduo selecionado não está presente no momento de coleta de dados por estar doente, ou não participa do estudo por falta de interesse.

Nos estudos baseados em coletas longitudinais, a perda de informação também pode ocorrer quando o participante não é acompanhado em todos os momentos da avaliação. Esse tipo de ausência, por sua vez, pode ser classificado como monotônico ou ausência com padrão intermitente (Little, 2024). A ausência monotônica acontece quando o indivíduo abandona o estudo, ou seja, deixa de participar da avaliação a partir de determinado momento, não retornando mais. No padrão intermitente, o participante se ausenta em determinado momento da avaliação, mas retorna no momento seguinte. Ainda, segundo Wærsted et al. (2018), a ausência também pode ser classificada como mista quando uma ausência intermitente é seguida de um padrão monotônico.

Segundo Cheema (2014), a ausência de dados em uma ou mais variáveis de interesse é constante nas pesquisas educacionais. O autor afirma que os bancos de dados relativos a avaliações educacionais em larga escala nos Estados Unidos, compostos de milhares de observações, raramente estão completos. A ocorrência de dados ausentes em avaliações educacionais em larga escala no Brasil também é comum, e essa perda de informação tem reflexo na produção científica da área. Na revisão realizada por Ferrão et al. (2020) foram identificados, no período de 2000 a

2018, 60 artigos nos quais foram utilizados dados incompletos da Prova Brasil;[1] no entanto, entre esses artigos, apenas 23 mencionam explicitamente a ocorrência de valores ausentes e em 18 são utilizados métodos para lidar com o problema.

As avaliações educacionais com base em levantamentos longitudinais também são muito afetadas por outro fator, a evasão escolar. Com isso, além da perda de informação referente à ausência nos momentos da avaliação, parcela considerável de estudantes pode abandonar os estudos por ter se evadido da escola. A evasão escolar é um processo cumulativo e ainda é elevada no Brasil, principalmente no ensino médio (Ministério da Educação [MEC], 2024). Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), a partir dos dados da Pesquisa Nacional por Amostra de Domicílios (Pnad) de 2023, verifica-se que, "no grupo etário de 14 a 29 anos, 9,0 milhões não completaram o ensino médio, seja por terem abandonado a escola antes do término desta etapa ou por nunca a terem frequentado" (Bello & Britto, 2024).

Nesse contexto, o objetivo do presente estudo é propor uma nova abordagem para o tratamento de dados ausentes, baseada no modelo de misturas de padrões.[2] Esse novo método é então comparado com o *listwise deletion* (LD) e com a imputação múltipla (IM). A comparação é realizada tendo como base os dados da avaliação longitudinal realizada no estado do Ceará, no período de 2009 a 2011, por meio do Sistema Permanente de Avaliação da Educação Básica (Spaece). O modelo de crescimento linear é o modelo principal usado nas comparações, tendo o desempenho dos estudantes em matemática como variável dependente. A seguir são apresentados conceitos relacionados aos dados ausentes e a descrição dos métodos utilizados nas comparações.

TIPOS DE DADOS AUSENTES

A teoria mais conhecida e utilizada para classificação de dados ausentes foi proposta por Rubin (1976). Essa classificação é utilizada no presente estudo e os mecanismos geradores de dados ausentes são apresentados utilizando o contexto de uma avaliação educacional longitudinal. Supõe-se que o desempenho acadêmico de n estudantes é monitorado em T anos por meio de testes de proficiência. Para um determinado estudante i, o conjunto de respostas planejadas (os escores nos testes de proficiência nos T anos) pode ser representado pelo vetor $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{it})'$. A ausência de valores pode ocorrer em parte dos testes inicialmente planejados no estudo, logo o vetor Y_i pode estar incompleto. A ausência ou presença de informação em Y_i é sinalizada em R_i, de tal forma que $R_{ij} = 1$ quando o estudante i não está presente na avaliação do ano j e $R_{ij} = 0$ quando esse estudante é avaliado no ano j.

1 Para mais informações, veja: http://portal.mec.gov.br/prova-brasil

2 O presente estudo faz parte da tese de doutorado do primeiro autor (Vinha, 2016).

De acordo com a informação contida em R_i, o vetor Y_i pode ser dividido em Y_{io} e Y_{ia}, que correspondem às respostas observadas e ausentes, respectivamente.

Os mecanismos geradores de dados ausentes, segundo a classificação proposta por Rubin (1976), podem então ser definidos de acordo com a distribuição dos indicadores de respostas (R_i), dado os valores de Y_{io}, Y_{ia} e X_i, onde X_i corresponde à matriz de covariáveis do indivíduo i. Em geral, nas avaliações educacionais, essas covariáveis são coletadas por meio de questionários contextuais e referem-se a características sociodemográficas e percepções acerca do ambiente escolar.

MCAR: Ausentes completamente ao acaso

Os valores ausentes são classificados como completamente ao acaso quando a ausência não está relacionada com as variáveis observadas no estudo ou com a variável que apresenta os valores faltantes. A ausência de valores do tipo *missing completely at random* (MCAR), observada no desempenho dos estudantes em determinado momento t, não teria então relação com o desempenho a ser avaliado, desempenho anterior ou qualquer outra variável presente no estudo. Assim, a probabilidade de ausência, dado os valores observados e não observados, é dada por:

$$P(R_i/Y_{io}, Y_{ia}, X_i) = P(R_i). \quad (1)$$

Os dados MCAR podem ser interpretados como uma amostra aleatória de estudantes retirada de um banco de dados completo. Logo, os resultados obtidos utilizando apenas os dados completos observados podem ser extrapolados para a população de interesse do estudo. Esse tipo de ausência não traz problemas adicionais na análise e interpretação de resultados, além da inevitável perda de precisão resultante da redução do tamanho da amostra. No entanto, a suposição de ausência completamente ao acaso parece pouco adequada no contexto de avaliações educacionais. Por exemplo, a ausência na avaliação pode ser causada pela evasão escolar, e a evasão geralmente está associada a condições socioeconômicas e desempenho do estudante (Ferreira, 2022).

É possível avaliar empiricamente a suposição de ausência do tipo MCAR. Uma vez que a perda de informação nessa situação acontece ao acaso, os indivíduos com valores ausentes não diferem daqueles com valores válidos, considerando as características avaliadas no estudo. Por exemplo, supondo que a ausência de valores relativos aos testes de proficiência seja completamente ao acaso, não existe diferença de perfil socioeconômico entre os estudantes ausentes e os presentes na avaliação. Dessa forma, essa suposição pode ser avaliada por meio de testes estatísticos tendo como hipótese nula a não diferença entre os grupos.

Se, por um lado, a suposição de ausência completamente ao acaso é muito forte e geralmente não se sustenta na prática, por outro, esse mecanismo pode ser

usado de forma planejada, com o intuito de diminuir os custos da avaliação. Por exemplo, em pesquisas longitudinais em larga escala, pode-se propor que em cada aplicação de testes de proficiência seja considerada apenas uma parcela dos estudantes, de tal forma que a seleção dos indivíduos seja completamente aleatória e independente das características analisadas (Fitzmaurice et al., 2008). Assim, é possível obter resultados muito próximos aos que seriam observados com uma amostra completa, mas com custo reduzido.

MAR: Ausentes ao acaso

Os dados ausentes do tipo *missing at random* (MAR) têm taxa de ocorrência relacionada às variáveis observadas no estudo (as covariáveis e a variável em questão em momentos anteriores), mas independe do valor não observado. Logo, a probabilidade de ausência é dada por:

$$P(R_i/Y_{io}, Y_{ia}, X_i) = P(R_i/Y_{io}, X_i). \quad (2)$$

Assim, as observações faltantes ocorrem com maior frequência em uma parcela da população. Por esse motivo, os dados completos não podem ser tratados como uma amostra aleatória da população em estudo. Médias e variâncias calculadas com base apenas nas observações completas podem gerar estimativas viesadas para os parâmetros populacionais.

Apesar do potencial de distorcer resultados, a suposição de que o mecanismo gerador de ausência é do tipo MAR traz um resultado importante relacionado à distribuição dos dados faltantes. Por exemplo, suponha que a ausência no escore de desempenho depende da renda familiar do estudante, mas para uma determinada faixa de renda a ausência é aleatória. Logo, em uma certa faixa de renda, não existe diferença entre a distribuição dos valores ausentes e a distribuição dos valores observados de desempenho. Por essa razão, a ausência do tipo MAR também é chamada de ignorável, uma vez que a análise pode gerar inferências válidas, quando utilizadas técnicas adequadas (Allison, 2001; Fitzmaurice et al., 2008). As técnicas baseadas na máxima verossimilhança e IM têm como suposição a ausência do tipo MAR e apresentam bons desempenhos nesses casos.

Apesar de razoável em muitas situações, a suposição de que os dados ausentes são do tipo MAR não parece adequada em outras (Enders, 2022). Por exemplo, estudantes insatisfeitos com a escola podem ter maior chance de não responder ao questionário sobre satisfação escolar, nesse caso a ausência depende dos valores não observados da variável em questão (Jeličić et al., 2009).

MNAR: Ausentes não ao acaso

Os dados faltantes são identificados como *missing not at random* (MNAR) quando a ocorrência está relacionada aos valores não observados, além de depender dos valores observados e das covariáveis presentes no estudo, logo, a probabilidade de ausência é dada por:

$$P(R_i/Y_{io}, Y_{ia}, X_i). \tag{3}$$

Uma vez que a ausência de dados não pode ser explicada somente pelos valores observados, esse tipo de ausência também é chamado de não ignorável, ou informativo (Graham, 2009).

Na avaliação educacional isso acontece quando existe uma taxa maior de valores faltantes nos testes de proficiência entre os alunos com menor desempenho no momento (mesmo depois de controladas outras variáveis). Trata-se do padrão de não resposta mais crítico e pode gerar sérias distorções nos resultados (Enders, 2022). É importante ressaltar que a diferenciação entre a ausência do tipo MAR e MNAR não pode ser realizada empiricamente, uma vez que depende dos dados não observados que, em geral, não podem ser recuperados. No exemplo, o desempenho não medido dada a ausência dos estudantes no dia da avaliação não pode ser coletado em outro momento. Esse fato representa um ponto crítico na escolha da abordagem a ser utilizada para análise, uma vez que depende de suposições não testáveis.

MÉTODOS PARA TRATAMENTO DE DADOS AUSENTES

Técnicas tradicionais de tratamento de dados ausentes – como a imputação pela média ou o LD, que consiste na retirada dos indivíduos com uma ou mais informações faltantes – ainda são muito utilizadas por pesquisadores (Davis et al., 2018; Jeličić et al., 2009; Rousseau et al., 2012). A imputação dos valores ausentes pela média da variável calculada com base nos valores observados na amostra, segundo Enders (2022), é o pior método de tratamento de dados ausentes por não ter embasamento teórico e não considerar as relações entre as variáveis. O uso desse procedimento produz distorções nas estimativas para qualquer tipo de ausência, uma vez que subestima a variabilidade dos dados, o que pode afetar os testes de significância, além de atenuar as medidas de associação. O LD apresenta como principal consequência a redução do poder dos testes estatísticos para qualquer mecanismo de ausência, principalmente quando o número de valores ausentes é elevado (Vinha & Laros, 2018); e, ainda, se os dados ausentes não são do tipo MCAR, ou seja, se a ausência está relacionada com as variáveis envolvidas no estudo, podem-se gerar estimativas viesadas para os parâmetros, sendo que a amostra com dados completos tende a reduzir a presença de indivíduos de certos subgrupos da população de interesse.

O desenvolvimento de métodos mais sofisticados para análise de dados incompletos tem sido intenso nas últimas décadas (Enders, 2023). Os métodos baseados na estimação por máxima verossimilhança e na IM têm recebido atenção especial por parte dos pesquisadores por apresentarem melhor desempenho que as técnicas tradicionais, principalmente quando os dados ausentes são do tipo MAR (Graham, 2009). Esses procedimentos baseiam-se na estimação dos modelos utilizando todas as informações disponíveis e considerando as estruturas de variâncias e covariâncias dos dados; com isso, quando a ausência é do tipo MAR, os valores ausentes podem ser estimados a partir das relações entre as variáveis, e, dessa forma, os parâmetros de interesse são obtidos com maior precisão (Enders, 2023). Ainda, na era da inteligência artificial, o uso de técnicas de aprendizado de máquina tem melhorado a *performance* de tratamento de dados ausentes em diferentes áreas (Alabadla et al., 2022; Ismail et al., 2022; Seu et al., 2022).

Neste estudo, os resultados do LD são comparados com os resultados da aplicação da IM e de um procedimento baseado na mistura de padrões (classe de modelos desenvolvidos para análise de dados do tipo MNAR). Essas metodologias são descritas a seguir.

Imputação múltipla

A imputação múltipla (IM) baseia-se na substituição dos valores ausentes realizada m vezes, gerando assim m versões plausíveis de bancos de dados completos (Little, 2024; Schafer & Graham, 2002). Essas m versões são então analisadas utilizando-se as técnicas convencionais, com isso estimativas pontuais e erros padrão são gerados a partir da combinação dos m resultados. Sob a suposição de que a ausência de dados é do tipo MAR, a IM gera estimativas não viesadas para os parâmetros de interesse; além disso, a incerteza relacionada à ausência de informação também é considerada nas análises (Little & Rubin, 2019).

Na IM os dados incompletos são tratados em uma etapa anterior ao ajuste do modelo principal de análise. A etapa de imputação é realizada por meio de dois passos: estimação do vetor de médias e da matriz de variâncias e covariâncias; e imputação dos valores ausentes, o que corresponde à retirada de uma amostra da distribuição *a posteriori* da matriz de variâncias e covariâncias e do vetor de médias. Esses passos são repetidos até que sejam gerados m bancos de dados com valores imputados (Enders, 2022). O número de bancos de dados gerados (m) e o de iterações são importantes para a eficiência da técnica.[3]

Na etapa de imputação deve-se considerar o uso de variáveis auxiliares. As variáveis auxiliares podem não fazer parte do modelo principal de análise, mas

3 Para mais detalhes, veja Graham et al. (2007).

estão associadas ao mecanismo gerador de valores ausentes ou à variável a ser tratada (Schafer & Graham, 2002). A inclusão dessas variáveis aumenta a chance de satisfazer a suposição de ausência do tipo MAR, e assim melhorar a estimação (Baraldi & Enders, 2010; Collins et al., 2001; Pigott, 2001). Além das variáveis auxiliares, as variáveis presentes no modelo principal devem ser consideradas na imputação.

Na etapa seguinte, os m conjuntos de dados imputados são analisados utilizando os métodos estatísticos usuais, gerando assim m estimativas pontuais e erros padrão para os parâmetros de interesse. Por fim, na última etapa, os resultados da etapa anterior são combinados, e assim são determinadas as estimativas finais.

Para a última etapa, Rubin (1987) propõe que as estimativas pontuais dos parâmetros sejam calculadas pela média aritmética dos valores obtidos nas m análises. Além das estimativas pontuais, os erros padrão também devem ser combinados; nesse caso, é proposta uma combinação que considera uma variação dentro das amostras e entre elas. A variação dentro (V_D) das amostras é dada por:

$$V_D = \frac{1}{m}\sum_{i=1}^{m} \widehat{SE}_i^2, \tag{4}$$

onde $\widehat{SE}_i^2$ é o erro padrão ao quadrado relativo ao parâmetro β na amostra i, e a variação entre as amostras (V_E) é calculada por:

$$V_E = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\beta}_i - \bar{\beta}\right)^2. \tag{5}$$

Assim, o erro padrão combinado corresponde à raiz quadrada da variação total (V_T), onde

$$V_T = V_D + V_E + V_E/n. \tag{6}$$

Logo, a variação total relacionada à estimação dos parâmetros é composta de uma parcela que corresponde à variação intrínseca dos dados (V_D), mais uma parcela que reflete a variação entre as amostras (V_E), resultado das diferentes imputações. Dessa forma, IM introduz uma incerteza relacionada à ausência de dados, o que não acontece quando a imputação simples pela regressão é utilizada (Allison, 2001).

Além da suposição de ausência do tipo MAR, a IM também requer uma suposição relacionada à distribuição dos dados; em geral, a distribuição normal multivariada é utilizada. Os resultados observados para esse método são muito semelhantes aos obtidos pelo método da máxima verossimilhança, especialmente quando o tamanho da amostra é grande (Collins et al., 2001).

Modelo de mistura de padrões

Os modelos de mistura de padrões foram desenvolvidos para a análise de dados com valores ausentes do tipo MNAR (Enders, 2022; Little, 2024; Little & Rubin, 2019). Esses modelos são baseados na estimação conjunta do modelo principal de análise e da propensão de ausência de dados.

Considere a seguinte distribuição conjunta de probabilidade:

$$f(Y_{it}, R_{it}/\theta, \phi), \tag{7}$$

onde Y_{it} é a variável resposta para o indivíduo i (o desempenho do estudante i em um determinado momento t da avaliação), e R_{it} é o indicador de ausência correspondente. O termo θ corresponde ao conjunto de parâmetros que descrevem a distribuição de Y (por exemplo, por meio de um modelo de crescimento linear), e ϕ contém o conjunto de parâmetros que descrevem a propensão de ausência de dados em Y (por exemplo, em um modelo de regressão logística). No modelo de mistura de padrões, essa distribuição conjunta é expressa pelo seguinte produto de distribuições:

$$f(Y_{it}, R_{it}/\theta, \phi) = f(Y_{it}/R_{it}, \theta).f(R_{it}/\phi), \tag{8}$$

onde $f(Y_{it}/R_{it}, \theta)$ é a distribuição condicional de Y_{it}, dado um particular valor de R_{it}, e $f(R_{it}/\phi)$ a distribuição de R_{it} dado o perfil do indivíduo. Por essa decomposição, o primeiro termo corresponde ao modelo principal de análise dos dados para grupos de indivíduos que compartilham o mesmo padrão de ausência, e o segundo termo corresponde ao modelo que descreve a incidência de valores ausentes desses grupos.

Na prática, a aplicação dessa classe de modelos consiste na estratificação da amostra em grupos de acordo com o padrão de ausência, e, para cada grupo, o modelo principal de análise é ajustado (Enders, 2022). As estimativas finais são obtidas utilizando os resultados de cada grupo, considerando o tamanho dos grupos na ponderação. Contudo o modelo não pode ser estimado devido à ausência de valores para alguns desses grupos; portanto, para que esse modelo possa ser usado, são necessárias suposições adicionais, também chamadas suposições de identificação. Por exemplo, considere a aplicação de um modelo de crescimento linear em um levantamento com três momentos distintos de avaliação, e dados ausentes com padrão monotônico. Os grupos formados de acordo com o padrão de ausência são: indivíduos presentes apenas no primeiro momento de avaliação; indivíduos presentes no primeiro e no segundo momento; e aqueles presentes em todos os momentos de coleta de dados. Para cada grupo, considerando que os tamanhos das amostras são adequados, o modelo de crescimento deve ser estimado. No entanto, como a variável

resposta foi observada apenas em um momento para o primeiro grupo, o modelo proposto não poderia ser estimado. Logo, uma suposição de identificação relacionada a desempenho desse grupo torna-se indispensável.

Uma estratégia para contornar o problema da estimação é a utilização da combinação de padrões de ausência. No exemplo acima, o pesquisador poderia supor que o padrão de desempenho dos indivíduos presentes apenas no primeiro momento é o mesmo padrão observado para os que estavam presentes nos dois primeiros momentos de avaliação. Essa estratégia, proposta por Hedeker e Gibbons (1997), faz com que a amostra seja dividida em dois grupos em que o modelo de crescimento linear possa ser ajustado.

Outra estratégia empregada para a estimação do modelo de mistura de padrões é a imposição de restrições. Nesse caso, o parâmetro que não pode ser estimado para um determinado grupo é substituído por estimativas obtidas em outros padrões. A imposição de restrição pode ser feita utilizando casos completos, casos vizinhos e casos disponíveis. Quando são considerados os casos completos, o parâmetro não identificado no grupo é substituído pela estimativa obtida para indivíduos com dados completos. De forma semelhante, a imposição pode ser feita utilizando os casos vizinhos, ou seja, os grupos com padrões de ausência mais próximos. Ou, ainda, usando todos os casos disponíveis, quando a estimativa aplicada na substituição é obtida a partir das estimativas de todos os grupos em que o parâmetro pode ser estimado. Mais detalhes e outras estratégias de identificação são apresentados em Demirtas e Schafer (2003).

Segundo Enders (2011), o modelo de misturas de padrões pode ser mais interessante do que outros presentes na literatura, como, por exemplo, os modelos de seleção. O modelo de misturas de padrões requer um tipo de suposição que obriga o pesquisador a explicitar a escolha; por outro lado, nos modelos de seleção, a suposição é imposta a uma distribuição, o que pode ser vago. Além disso, o modelo de mistura de padrões possibilita estudos de sensibilidade através da utilização de diferentes restrições.

Procedimento proposto

Antes de escolher a abordagem a ser utilizada na análise dos dados, o pesquisador deve avaliar as suposições mais adequadas em relação ao mecanismo gerador da ausência (Pigott, 2001). Detalhes sobre o processo de coleta de dados e o conhecimento da área de pesquisa são fundamentais nessa avaliação. Os dados a serem analisados no presente estudo referem-se ao acompanhamento dos estudantes do ensino médio no estado do Ceará; deve-se, portanto, considerar, na escolha da abordagem, que a trajetória desses alunos é influenciada pelo contexto socioeconômico.

A evasão escolar no Brasil não é um problema recente e é resultado da ação de diversos fatores (Ferreira, 2022). Leon e Menezes (2002) analisaram reprovação, avanço e evasão escolar por meio dos dados da Pesquisa Mensal de Emprego (PME), entre 1985 e 1997. Esse estudo indicou que idade, sexo, renda, trabalho e composição familiar são importantes para explicar a trajetória escolar no ensino médio. Com base nos dados da Pnad de 2004 e 2006, Neri (2009) verificou que a falta de interesse pela escola e a necessidade de renda são os fatores mais importantes para explicar a evasão escolar entre os indivíduos de 15 a 17 anos. Soares et al. (2015) identificaram que o abandono escolar no ensino médio em Minas Gerais é afetado por composição familiar, defasagem idade/série, trabalho, condição socioeconômica, gravidez e dificuldade nas disciplinas.

Especificamente para os estudantes do ensino médio do estado do Ceará, a partir dos dados do Spaece, Shirasu (2014) verificou que a chance de evasão entre os indivíduos com reprovações é o dobro da chance dos indivíduos sem reprovação. Além disso, a autora identificou que as variáveis sexo, distorção idade/série e escolaridade dos pais são importantes preditoras tanto da repetência quanto da evasão escolar.

Com isso, para o presente estudo, optou-se pela elaboração de um procedimento que tenha como suposição a ausência do tipo MNAR. A escolha da classe de mistura de padrões neste caso é motivada pela forte relação entre a repetência e a evasão escolar, além da relação entre a evasão e outras características dos estudantes, como renda, escolaridade da mãe, trabalho do estudante fora de casa e o próprio desempenho acadêmico. Esse procedimento propõe a mistura de padrões entre repetentes e ausentes. A suposição principal é que os estudantes com valores ausentes no desempenho acadêmico teriam evolução semelhante aos estudantes com reprovação no ensino médio. Essa suposição baseia-se na relação entre a reprovação e a evasão, ambas relacionadas a características socioeconômicas, mas também resultantes da falta de motivação e do baixo desempenho dos estudantes nas disciplinas.

Verificou-se também que, dentro dos grupos formados pela combinação de padrão de ausência e repetência, ainda existia grande heterogeneidade no perfil dos estudantes. Foi proposta então, para cada grupo, uma estratificação baseada nas variáveis que estão relacionadas com a ausência de dados. Verificou-se que idade, turno, escolaridade da mãe, trabalho e desempenho em língua portuguesa são importantes preditores da ausência na variável desempenho em matemática. Assim, visando a criar estratos mais homogêneos de alunos, foi proposta uma subdivisão dos grupos de acordo com o escore de propensão a ausência a partir dessas variáveis.

Por fim, o procedimento utiliza a combinação dos resultados, como na IM, para a geração das estimativas pontuais e erros padrão. Nesse caso, uma modificação

foi proposta no procedimento: a etapa de imputação foi realizada separadamente para cada estrato proposto, de tal forma que ocorra a mistura de padrões de acordo com as suposições acima.

MÉTODO

Dados

Os dados utilizados no presente estudo referem-se à avaliação dos estudantes do ensino médio do estado do Ceará realizada pelo Spaece.[4] Nesse sistema, os estudantes são avaliados nos três anos do ciclo por meio de testes de proficiência em matemática e língua portuguesa e respondem a questionários contextuais. As proficiências são estimadas por meio da teoria da resposta ao item e expressas na escala utilizada pelo Sistema Nacional de Avaliação da Educação Básica (Saeb), com média 250 e desvio padrão 50 (Secretaria da Educação, 2011). O questionário contextual respondido pelos estudantes abrange características sociodemográficas, rotinas de estudo, percepções sobre a escola e perspectivas com relação ao futuro.

Milhares de estudantes são avaliados anualmente pelo sistema, porém grande parte da informação é perdida. Para o presente estudo foi selecionada uma amostra de alunos que cursavam o primeiro ano do ensino médio em 2009. Além das informações de 2009, foram consideradas as proficiências avaliadas nos dois anos seguintes, 2010 e 2011, o que corresponde ao segundo e terceiro ano do ciclo para os estudantes sem reprovação, e a proficiência avaliada em 2008, quando esses estudantes estavam no último ano do ensino fundamental.

Os arquivos de dados referentes a essas avaliações não têm uma identificação comum para os estudantes de um ano para outro. Como consequência, esses arquivos não podem ser interligados diretamente por um identificador numérico comum; nesse caso, o nome dos estudantes foi utilizado como indexador. Após o processo de interligação dos arquivos, foi consolidada uma base composta das informações relativas a 8.681 estudantes.

Da forma como a amostra foi selecionada, todos os estudantes estavam presentes na avaliação de 2009, logo todos têm o registro dos testes de proficiência e das informações referentes ao questionário contextual naquele ano. As proficiências relativas a 2010 e 2011 estão incompletas, como apresentado na Tabela 1. Observa-se que 450 estudantes apresentam padrão intermitente de ausência, uma vez que não estavam presentes na avaliação de 2010, mas estavam em 2011, logo esses indivíduos não se evadiram da escola e a ausência deveu-se a outro motivo desconhecido. Entre os que estavam presentes somente em 2009, ou em 2009 e 2010, com base nos

4 Para mais informações, veja: https://www.seduc.ce.gov.br/spaece/

dados disponibilizados, não é possível identificar se eles apenas faltaram nos dias das avaliações ou se a ausência é monotônica causada pela evasão escolar. Nesse caso, seria necessária a utilização de outras fontes de informação para identificar se os estudantes ausentes estavam matriculados na escola ou não.

TABELA 1
Padrões de ausência de dados

PADRÃO	2009	2010	2011	NÚMERO DE ESTUDANTES
0	P	P	P	6.447
1	P	A	P	450
2	P	P	A	652
3	P	A	A	1.132
Total				8.681

Fonte: Elaboração dos autores.
Nota: P = presente; A = ausente.

Além das proficiências, algumas variáveis relacionadas aos alunos foram utilizadas no presente estudo (Tabela 2). Essas variáveis foram coletadas por meio do questionário respondido pelos estudantes em 2009.

TABELA 2
Variáveis utilizadas no estudo

VARIÁVEL	DESCRIÇÃO/CODIFICAÇÃO
MAT09	Desempenho em matemática em 2009
LP09	Desempenho em língua portuguesa em 2009
MAT10	Desempenho em matemática em 2010
MAT11	Desempenho em matemática em 2011
MAT08	Desempenho em matemática em 2008, no último ano do ensino fundamental
Sexo	0: feminino; 1: masculino
Etnia	Etnia autodeclarada: branco, pardo, negro, amarelo e indígena
Turno	Turno em que o estudante frequentava as aulas: manhã, tarde ou noite
Idade	Idade em 2009 (em anos)
Escolaridade da mãe	A escolaridade da mãe é classificada em: • nunca estudou • 1ª a 4ª série do ensino fundamental • 5ª a 8ª série do ensino fundamental • 1ª a 3ª série do ensino médio • ensino superior • não sabe
Reprovações	Número de reprovações no ensino fundamental: 0: nunca repetiu; 1: uma reprovação; 2: duas reprovações; 3: três ou mais repetências

(continua)

(continuação)

VARIÁVEL	DESCRIÇÃO/CODIFICAÇÃO
Superior	Pretensão de ingresso no ensino superior. Assume o valor 1 se o aluno afirma que pretende ingressar no ensino superior e 0 se tem outros planos
Gosta de matemática	Assume o valor 1 se o aluno responde que matemática é a disciplina preferida e 0 se prefere outra
Dever	Assume o valor 1 se o aluno faz as tarefas de casa sozinho e 0 se não faz sozinho ou é auxiliado por outra pessoa
Repetência no ensino médio	Número de vezes que o aluno repetiu um ano escolar no ensino médio. Obtido a partir do registro da etapa cursada em cada ano: • 0: nunca repetiu • 1: uma repetência • 2: duas repetências
Trabalho	O estudante trabalha fora de casa. Assume o valor 1 se o aluno afirma que trabalha fora de casa e 0 se tem outras atividades

Fonte: Elaboração dos autores.

Procedimentos

A comparação entre os três procedimentos de tratamento de dados ausentes apresentada neste estudo baseia-se na estimação do modelo de crescimento linear que pertence ao grupo geral de análise de regressão multinível (Raudenbush & Bryk, 2002). Esse modelo é apresentado considerando uma estrutura em dois níveis: o primeiro nível corresponde às observações "dentro" de cada indivíduo e o segundo é o nível dos indivíduos. Dados o padrão observado de crescimento do desempenho e o número reduzido de observações para cada estudante, o componente temporal utilizado é descrito por uma função do primeiro grau. Nesse caso, a equação que descreve o primeiro nível é dada por:

$$Y_{it} = \pi_{0i} + \pi_{1i}(ANO)_t + \varepsilon_{it}, \tag{9}$$

onde Y_{it} é a proficiência em matemática do aluno i, no ano t; π_{0i} é a proficiência esperada do aluno i no início do ensino médio; π_{1i} é a taxa de aprendizado do aluno i em um ano acadêmico; e ε_{it} é a parcela aleatória da proficiência não explicada pelo modelo.

A proficiência em matemática no final do ensino fundamental (MAT08), a pretensão de ingresso no ensino superior (Superior) e se o estudante está trabalhando fora de casa (Trabalho) foram utilizadas como variáveis independentes no nível dos alunos. As equações para esse nível são:

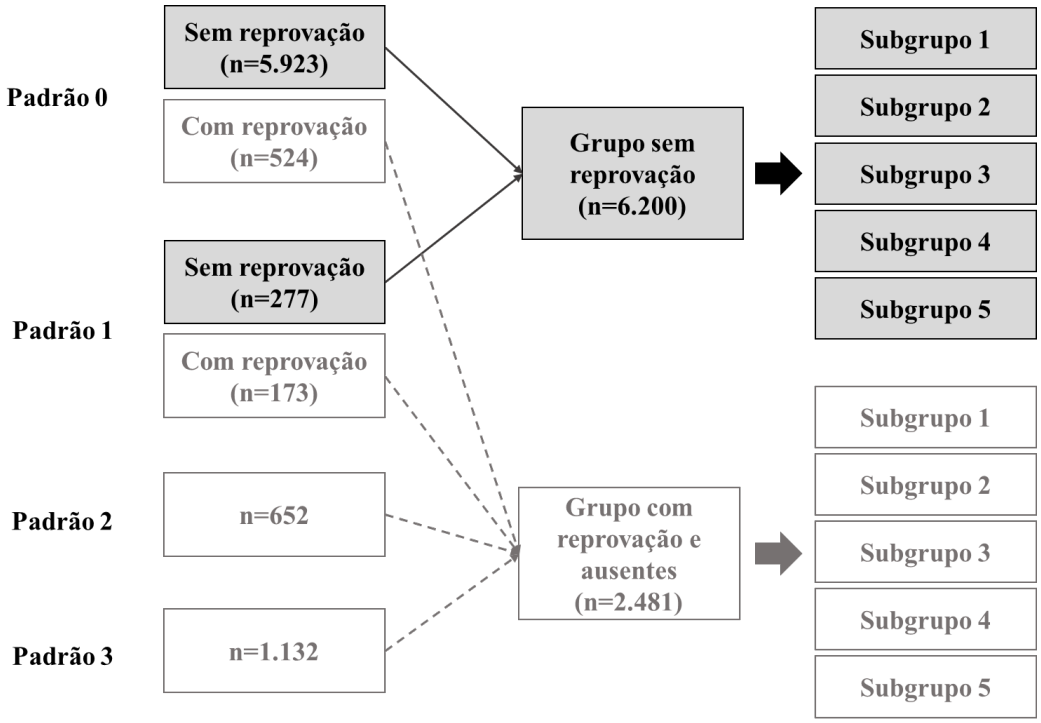
$$\pi_{0i} = \beta_{00} + \beta_{01}(Trabalho)_i + \beta_{02}(Superior)_i + \beta_{03}(MAT08)_i + r_{0i}, \tag{10}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(Trabalho)_i + r_{1i.} \tag{11}$$

Por essa formulação, as variáveis MAT08 e Superior têm influência na proficiência inicial e a variável Trabalho é importante tanto para a proficiência inicial quanto para a taxa de crescimento. Nesse caso, o coeficiente β_{01} representa a diferença no desempenho no início do ensino médio entre o grupo de estudantes que trabalham e os que não trabalham. O coeficiente β_{11} quantifica a mudança na taxa anual de crescimento para os estudantes que trabalham em relação aos demais. Os termos r_{0i} e r_{1i} correspondem aos componentes aleatórios da proficiência esperada e da taxa de aprendizado.

O procedimento de mistura de padrões proposto baseia-se na divisão dos dados para a estimação com dados ausentes. Nesse caso, o banco de dados foi dividido de acordo com o padrão de ausência, a ocorrência de reprovação e o escore de propensão de ausência no final do estudo. Inicialmente os estudantes foram divididos em dois grupos: os estudantes sem reprovação no ensino médio e que apresentam padrão 0 e 1 de ausência (Tabela 1); e todos os outros com uma ou duas reprovações no ensino médio ou que apresentavam padrão 2 e 3 de ausência. Em seguida, uma nova divisão é proposta e esses grupos são então divididos de acordo com o escore de propensão. Para isso foram calculados os decis do escore de propensão dentro de cada grupo e os estudantes foram divididos em 5 subgrupos, tendo como limites o 2º, 4º, 6º e 8º decis. A Figura 1 apresenta esquematicamente essa divisão.

FIGURA 1
Divisão dos estudantes no procedimento de mistura de padrões



Fonte: Elaboração dos autores.

Análise de dados

Todas as análises foram realizadas utilizando o *software* estatístico *Statistical Analysis System* (SAS) versão 9.4. Os comandos utilizados em cada procedimento foram:

- *Listwise deletion* (LD): PROC MIXED foi utilizada para o ajuste do modelo de crescimento linear, considerando apenas as observações completas (Padrão 0).
- Imputação múltipla (IMD): a PROC MI foi empregada na etapa de imputação com as covariáveis MAT08, LP09, Sexo, Turno, Idade, Escolaridade da mãe, Superior, Gosta de matemática, Dever, Repetência no ensino fundamental e Trabalho. Foram gerados cem bancos de dados imputados, com quinhentas iterações iniciais antes da retirada do primeiro banco e duzentas iterações entre as retiradas subsequentes. A PROC MIXED foi utilizada na etapa de estimação, o modelo de crescimento linear proposto foi estimado com base nas cem amostras geradas na etapa anterior. A PROC MIANALYZE foi usada para gerar as estimativas finais a partir dos resultados das cem estimativas obtidas na etapa anterior segundo o método proposto por Rubin (1987).
- Mistura de padrões: as funções MI, MIXED, MIANALYZE e LOGISTIC foram utilizadas. No entanto, nesse caso, a etapa de imputação é realizada separadamente em cada estrato formado pela combinação de grupo, reprovação e escore de propensão descritos anteriormente. Os escores usados foram gerados pela regressão logística, tendo como variável resposta a ausência no último ano do estudo e como independentes as variáveis MAT09, LP09, Turno, Escolaridade da mãe, Idade, Sexo, Repetência no ensino fundamental, Trabalho e Superior.

RESULTADOS

A Tabela 3 apresenta o perfil dos estudantes por meio da distribuição das variáveis coletadas em 2009. Para os 8.681 estudantes selecionados na amostra, observou-se idade média de 15,83 anos (DP = 1,06), 54,12% eram estudantes do sexo feminino e 39,73% estudavam no turno da manhã. Vale destacar que o registro da reprovação no ensino médio foi possível apenas para os estudantes presentes nos anos de 2010 ou 2011 ou em ambos, nesse caso o percentual foi obtido com base nas respostas de 7.549 estudantes. Entre esses estudantes, 10,56% repetiram uma ou duas vezes no ensino médio.

TABELA 3
Perfil da amostra

VARIÁVEL	%
Sexo	
Feminino	54,12
Masculino	45,88
Turno	
Manhã	39,73
Tarde	35,20
Noite	25,07
Etnia	
Branco	19,38
Pardo	58,05
Negro	12,17
Amarelo	6,49
Indígena	3,91
Escolaridade da mãe	
Nunca estudou/não sabe	22,16
1ª a 4ª do EF	31,71
5ª a 8ª do EF	22,23
1ª a 3ª do EM	17,53
Superior	6,36
Superior	
Não	59,27
Sim	40,73

VARIÁVEL	%
A disciplina preferida é matemática	
Não	81,53
Sim	18,47
Faz o dever de casa sozinho	
Não	36,30
Sim	63,70
Trabalha fora de casa	
Não	84,60
Sim	15,40
Número de reprovações no EF	
0	62,38
1	24,08
2	10,14
3 ou mais	3,41
0	62,38
Número de reprovações no EM*	
0	89,35
1	9,43
2	1,22

Fonte: Elaboração dos autores.

Nota: EF = ensino fundamental; EM = ensino médio; Superior = pretensão do aluno de ingressar no ensino superior.

* Os percentuais referem-se ao total de 7.549 estudantes.

Considerando todos os estudantes envolvidos no estudo, verifica-se que a proficiência média em matemática em 2009 é de 242,41 pontos, com desvio padrão de 45,76 pontos (Tabela 4). Nota-se também que a proficiência média varia de acordo com o padrão de ausência de dados durante o ciclo, com diferença de 20,45 pontos entre as médias dos estudantes presentes em todo o acompanhamento (padrão 0) e daqueles presentes apenas em 2009 (padrão 3).

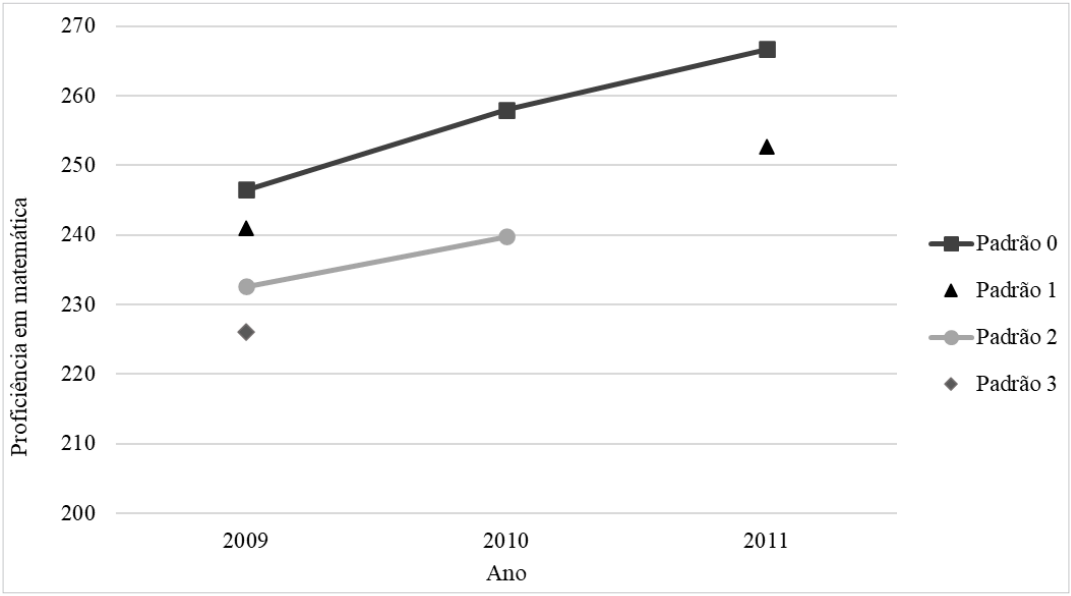
TABELA 4
Perfil da amostra

PROFICIÊNCIA EM MATEMÁTICA EM 2009	PADRÃO 0	PADRÃO 1	PADRÃO 2	PADRÃO 3
Média	246,38	240,73	232,9	225,93
Desvio padrão	46,32	43,59	44,1	39,46

Fonte: Elaboração dos autores.

As trajetórias dos alunos de acordo com o padrão de ausência e a reprovação no ensino médio foram analisadas, considerando apenas os valores válidos da amostra. Pode-se observar, pela Figura 2, que os alunos presentes nas três avaliações realizadas no ensino médio têm proficiência média maior que os demais grupos e apresentam maior taxa de crescimento no período.

FIGURA 2
Proficiência média em matemática dos estudantes dos diferentes padrões de ausência



Fonte: Elaboração dos autores.

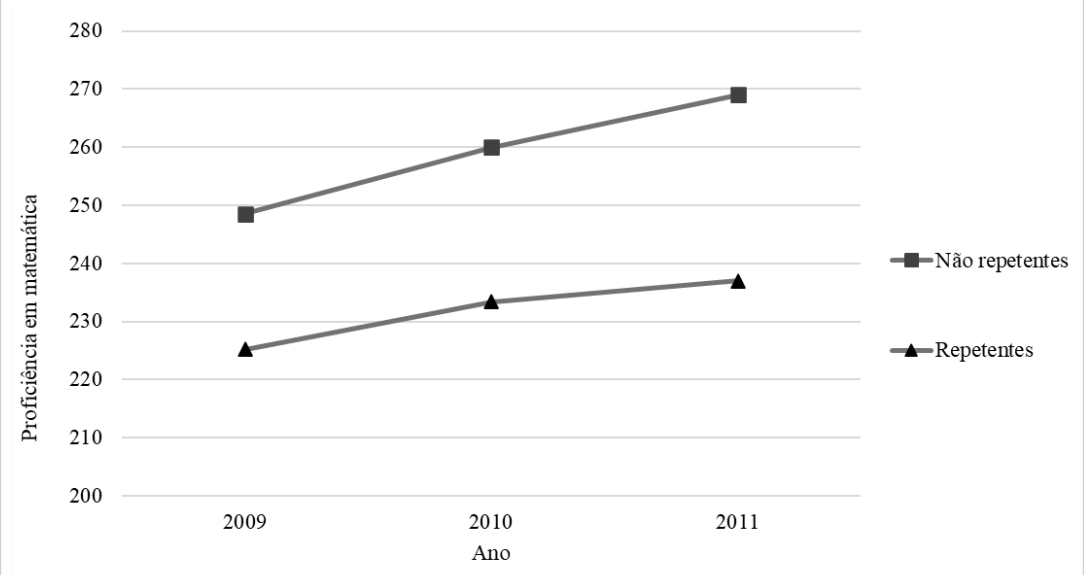
A Figura 2 sinaliza alguns pontos importantes que foram considerados no tratamento dos dados. Os indivíduos sem valores ausentes têm, em média, melhor desempenho em matemática, seguidos pelos estudantes que não foram avaliados apenas em 2010 (padrão 1). Ainda, observa-se que os estudantes ausentes em 2011 (padrão 2) e os estudantes avaliados apenas em 2009 (padrão 3) mostram um desempenho médio em matemática inferior.

Esses resultados ressaltam a relação entre a ausência nas avaliações e a evasão escolar com o desempenho acadêmico. Por exemplo, a diferença de trajetória dos estudantes do padrão 1 e do padrão 0 pode indicar a relação entre desempenho

e ausência na avaliação em 2010 (por abandono naquele ano ou por outro motivo). Ainda, a diferença da proficiência média em 2009 dos estudantes do padrão 3, com ausência em dois momentos ou que se evadiram da escola, e os estudantes do padrão 1, apenas ausentes em 2010, sugere a relação entre evasão e desempenho.

A reprovação no ensino médio também é importante na evolução da proficiência em matemática dos estudantes no período analisado. Pela Figura 3, verifica-se que os estudantes reprovados em pelo menos um ano escolar durante o ensino médio têm desempenho inferior aos que não reprovaram. Pode-se observar que os repetentes têm proficiências médias menores e também uma evolução menos acentuada.

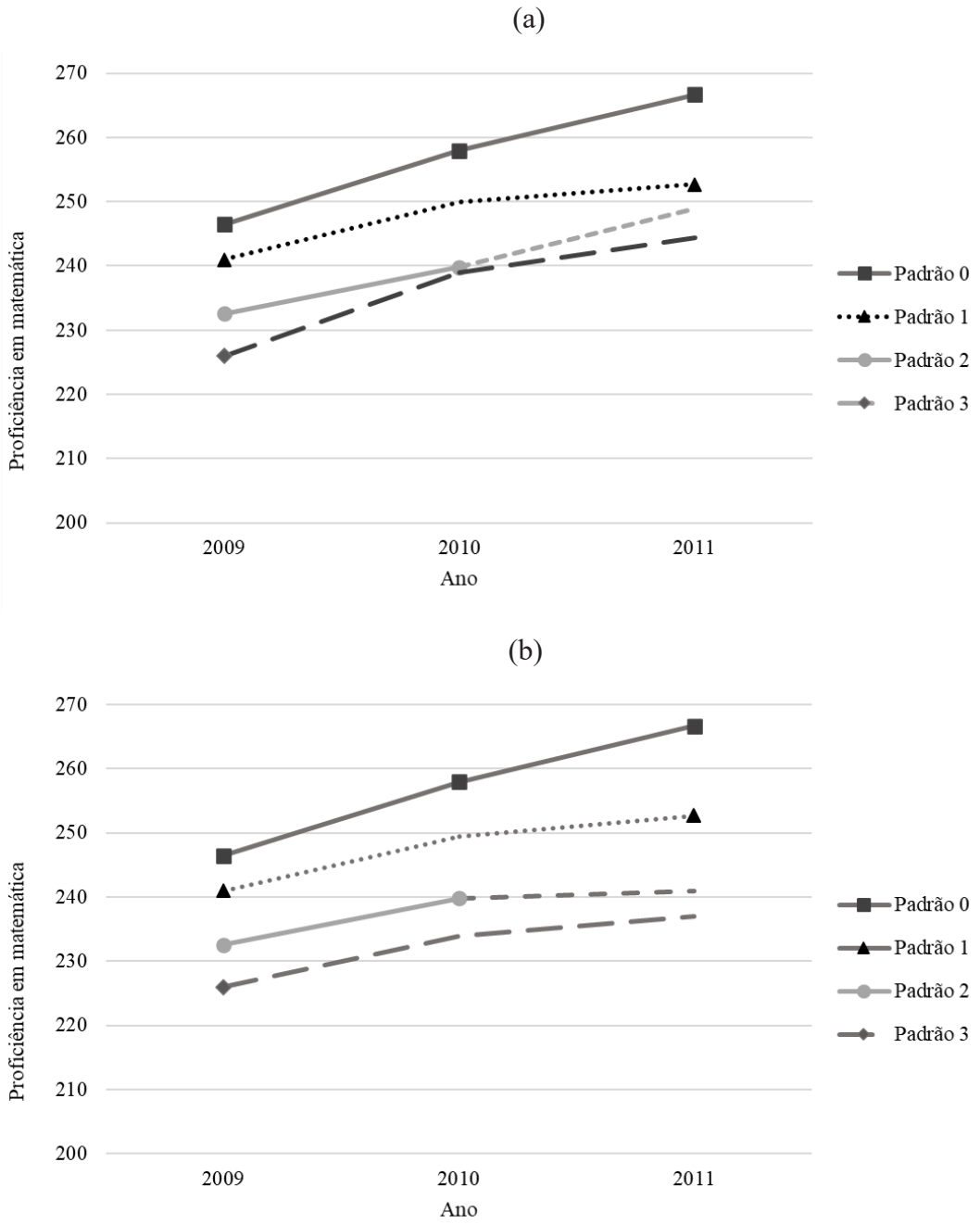
FIGURA 3
Proficiência média dos estudantes sem reprovação e com reprovação (n = 7.549)



Fonte: Elaboração dos autores.

Em seguida foram realizadas as imputações dos dados de acordo com os procedimentos de IM e MP. A Figura 4 apresenta a estimação do desempenho para os estudantes ausentes utilizando esses procedimentos – as curvas em pontilhado correspondem às trajetórias estimadas. Pode-se observar que, para os padrões 2 e 3 de ausência, a trajetória estimada usando o procedimento IM (Figura 4a) apresenta estimativas mais elevadas para o desempenho se comparadas com as estimativas do procedimento MP (Figura 4b). Esse resultado é consequência da abordagem adotada, uma vez que o procedimento MP supõe que os indivíduos com dados ausentes apresentam trajetória semelhante aos repetentes.

FIGURA 4
Proficiência média observada e estimada para os diferentes padrões de ausência: (a) procedimento IM e (b) procedimento MP



Fonte: Elaboração dos autores.

Por fim, na Tabela 5 são apresentados os resultados do ajuste do modelo de crescimento linear, utilizando os três procedimentos. De forma geral, o LD apresenta estimativas diferentes das obtidas pelos procedimentos IM e MP, e, entre esses dois últimos, as estimativas são mais próximas, principalmente para os parâmetros fixos do modelo.

TABELA 5
Comparação dos três procedimentos usando o modelo de crescimento linear

	LD			IM			MP		
Efeito fixo	Efeito	EP	Razão T	Efeito	EP	Razão T	Efeito	EP	Razão T
Intercepto	234,62	0,56	–	233,36	0,35	–	234,14	0,38	–
Ano	10,58	0,28	37,76	9,57	0,21	44,63	8,85	0,22	39,64
MAT08	0,66	0,01	67,67	0,65	0,01	86,82	0,64	0,01	86,09
Superior	7,47	0,85	8,75	9,29	0,66	14,02	9,24	0,66	13,92
Trabalho	-3,74	1,54	-2,45	-5,02	0,98	-5,13	-5,70	0,99	-5,76
Trabalho*Ano	-0,79	0,91	-0,86	-1,84	0,60	-3,11	-1,80	0,61	-2,91
EFEITO ALEATÓRIO									
Intercepto		492,39			464,30			440,10	
Ano		23,44			29,09			33,37	
Erro		739,91			740,18			795,46	
N		5.162*			8.681			8.681	

Fonte: Elaboração dos autores.

Nota: EP = erro padrão da média; Ano = taxa de crescimento por ano.

* Os estudantes com dados completos correspondem ao total do padrão 0, excluindo aqueles com valores faltantes nas variáveis Superior, Trabalho ou MAT08.

Na primeira parte da Tabela 5 – referente ao efeito fixo –, o valor do intercepto representa o valor médio da proficiência em matemática no início do ensino médio. Os procedimentos IM e MP apresentam intercepto menor, mas as diferenças observadas são relativamente pequenas, com variação inferior a 0,1% em relação ao valor 234,62 observado com o LD. Pode-se verificar que os procedimentos IM e MP apresentam efeitos mais acentuados para as variáveis relativas à intenção de ingresso no ensino superior e ao trabalho do estudante fora de casa. O coeficiente estimado para a variável Superior é aproximadamente 24% maior em relação ao resultado de LD, com estimativas iguais a 9,29 e 9,24. Para a variável Trabalho observa-se que o procedimento IM apresenta estimativa igual a -5,02, um efeito 34,2% mais acentuado em relação ao obtido em LD (-3,47); no procedimento MP o efeito estimado para essa variável foi de -5,70, um aumento em módulo de 52,4% em relação à estimativa em LD. Nota-se também que a interação entre Trabalho e Ano não é significativa no procedimento LD, mas tem efeito significativo quando os procedimentos IM e MP são utilizados, com estimativas iguais a -1,84 e -1,80, respectivamente.

Os resultados também ressaltam a influência da abordagem na estimativa do componente relacionado à taxa de crescimento da proficiência dos estudantes (Tabela 5). No procedimento LD verifica-se que a taxa de crescimento é de 10,58 pontos por ano, para a IM a taxa cai para 9,57 pontos e na MP o valor observado é 8,85 pontos.

Por fim, além dos efeitos fixos, no modelo de crescimento linear utilizado são estimadas as variâncias dos efeitos aleatórios da proficiência no início do período, taxa de crescimento e da parcela da proficiência não explicada pelo modelo (valores relativos a Intercepto, Ano e Erro, da segunda parte da Tabela 5, respectivamente). Verifica-se que no procedimento LD, a proficiência no início do período apresenta maior variação entre os estudantes, e, por outro lado, o procedimento MP estima um modelo que apresenta maior variação das taxas de aprendizado entre os estudantes.

DISCUSSÃO

Diversos métodos para o tratamento de dados ausentes são encontrados na literatura (Enders, 2023). A escolha do método adequado em cada situação é fundamental para mitigar os possíveis efeitos negativos da ausência de informação. Nesse sentido, a qualidade dos resultados obtidos a partir do uso de um determinado método depende diretamente da adequação das suposições relacionadas ao tipo de ausência. O objetivo do presente estudo foi apresentar uma comparação de três métodos baseados em suposições diferentes acerca do mecanismo gerador de ausência, tendo como base os dados longitudinais de uma avaliação educacional.

Os três procedimentos apresentaram estimativas distintas para os coeficientes do modelo de crescimento linear (Tabela 4). Essas diferenças podem ser explicadas pelas suposições implícitas no uso de cada abordagem. A utilização do LD é apropriada quando os dados ausentes são completamente ao acaso (MCAR), o que, nesse contexto, seria ausência não relacionada com desempenho e perfil dos estudantes. Uma vez que a suposição do procedimento LD é fortemente violada, sua utilização resulta em uma superestimação da taxa média de aprendizado dos estudantes. Além disso, verifica-se que os efeitos das variáveis relacionadas a trabalho fora de casa e intenção de ingresso no ensino superior são menores, o que também pode ser explicado pela retirada de parte não aleatória da amostra, deixando os melhores estudantes na amostra com informações completas.

Com a utilização do procedimento IM assume-se que a ausência seria explicada pelo perfil dos estudantes e pelo desempenho observado nos períodos anteriores (ausência do tipo MAR). Com isso, o tratamento dos valores faltantes considera as características dos estudantes ausentes, assim os desempenhos estimados para esses indivíduos são mais próximos aos observados para indivíduos com perfis semelhantes. Essa mudança no tratamento dos dados dos estudantes ausentes provavelmente é responsável pela redução da taxa média de aprendizado e pelo aumento do efeito do trabalho fora de casa e da intenção de ingresso no ensino superior em comparação ao procedimento LD, além do efeito significativo da interação entre o trabalho fora de casa e o ano da avaliação.

Por fim, o procedimento MP foi utilizado supondo que a ausência é do tipo MNAR. Nesse caso, a ausência não está apenas relacionada ao perfil dos estudantes e ao desempenho anterior, mas depende do valor que seria observado. No método proposto com a mistura de padrões, assumiu-se que o desempenho dos estudantes ausentes seria semelhante ao desempenho dos estudantes com uma ou mais reprovações. Como resultado, em relação à IM, verifica-se que a taxa média de aprendizado estimada é menor e o impacto da variável relativa ao trabalho fora de casa é maior.

A suposição de ausência completamente ao acaso é inadequada para os dados da avaliação do ensino médio cearense (Shirasu, 2014). Os indivíduos ausentes ou que abandonaram o ensino médio no Ceará têm perfis diferentes em termos de características socioeconômicas e do próprio desempenho escolar. Logo, os resultados obtidos a partir do procedimento LD devem ser interpretados com cautela. Ainda, baseado nos resultados do presente estudo, pode-se afirmar que o uso do procedimento LD não é recomendado para a análise de dados incompletos em avaliações educacionais longitudinais no Brasil.

Como mencionado, não é possível avaliar empiricamente se os dados ausentes são do tipo MAR ou MNAR, com isso não é possível afirmar com certeza qual dos procedimentos (IM ou MP) é mais adequado para os dados. No entanto é razoável supor que a ausência depende do desempenho a ser avaliado, principalmente nos casos em que a ausência é causada pela evasão escolar. Além disso, pela Figura 4, verifica-se que o tratamento dos dados pela IM estima trajetórias não esperadas para os grupos com ausentes, com crescimento semelhante ao observado para os estudantes com dados completos. As trajetórias geradas pelo procedimento MP são mais coerentes nesse contexto, os indivíduos com valores ausentes têm desempenho menor e taxa de aprendizado menor que os estudantes presentes em todo o estudo (Figura 4). Por exemplo, entre os anos de 2009 e 2010, o grupo 0 apresenta um aumento de 11,5 pontos na proficiência média, e, para o mesmo período, pelo IM, o grupo 3 tem proficiência média estimada de 239,1 pontos, resultando em um aumento maior, de aproximadamente 13 pontos. Para esse mesmo grupo 3, o procedimento MP estima proficiência média de 234,2 pontos em 2010, resultando em um aumento de cerca de 8 pontos em relação a 2009.

Em geral, os dados das avaliações educacionais no Brasil são incompletos, podem apresentar diferentes padrões de ausência, e o presente estudo contribui com a discussão da importância da escolha da abordagem a ser utilizada nas análises. O estudo de Ferrão et al. (2020) demonstra que a perda de informação na Prova Brasil pode ocorrer por ausência dos estudantes no dia da avaliação, estudantes presentes que não respondem ao questionário contextual e/ou o teste de proficiência, e ressaltam o uso da IM. Ferrão e Prata (2019) e Vinha e Laros (2018) discutem a escolha da

abordagem utilizada com simulações de padrões de ausência a partir de dados reais e relatam a importância do uso dos procedimentos baseados nas relações entre as variáveis (métodos da máxima verossimilhança e IM), além do uso de variáveis auxiliares. As discussões apresentadas neste estudo estendem essa discussão, abordando os padrões de ausência nas avaliações com dados longitudinais.

Além da comparação das trajetórias no período analisado, os perfis dos estudantes dos grupos formados de acordo com o padrão de ausência também foram comparados a partir dos dados de 2009. De forma geral, verificaram-se maiores percentuais de estudantes do sexo feminino, mais novos, que não trabalham e que pretendem ingressar no ensino superior entre os que estavam nos três momentos de avaliação (Padrão 0) com maiores diferenças, principalmente, em relação ao grupo presente apenas em 2009. Essa comparação está alinhada com os resultados de Shirasu (2014) e demonstra a relação intrínseca entre desempenho, evasão e outras características dos estudantes.

O presente estudo mostra algumas limitações impostas pelos dados utilizados. Os bancos de dados disponibilizados contêm informações relativas a milhares de estudantes avaliados anualmente no ensino médio do estado do Ceará, porém, dada a dificuldade de interligação dos arquivos, apenas uma parcela dessa informação foi utilizada. Ainda, o procedimento de junção dos dados utilizado é passível de erro, uma vez que a avaliação foi realizada pela similaridade de nomes. Além disso, com os dados disponibilizados, não é possível verificar se os indivíduos não avaliados em um ano estavam matriculados ou não; para obter essa informação, seria necessária a utilização do Censo Escolar ou alguma outra fonte.

O procedimento MP proposto foi desenvolvido tendo como fundamentação os resultados apresentados por outros autores (Leon & Menezes, 2002; Neri, 2009; Shirasu, 2014; Silva, 2013; Simões, 2014; Soares et al., 2015), que identificaram a relação entre o abandono e a repetência com o desempenho escolar e o perfil socioeconômico dos estudantes. A elaboração de um modelo baseado na mistura de padrões foi motivada pela busca de uma abordagem em que a suposição de identificação fosse clara e expressa em termos do contexto. O modelo proposto neste estudo tem como possíveis limitações a ausência de boas medidas relacionadas a variáveis apontadas por outros autores como fatores associados à evasão escolar, tais como nível socioeconômico, composição familiar, motivação e percepção da importância dos estudos na vida profissional. Vale também ressaltar que outras abordagens para a análise de dados com valores ausentes do tipo MNAR poderiam ser aplicadas aos dados, como os modelos de seleção e procedimentos baseados na reponderação da amostra (Fitzmaurice et al., 2008). Além disso, outros modelos baseados na mistura de padrões poderiam ser empregados, com a utilização de diferentes restrições de identificação.

Por fim, o uso de dados longitudinais pode trazer resultados interessantes acerca do aprendizado dos estudantes e é considerado por alguns autores como o "padrão-ouro" na avaliação de sistemas educacionais (Franco, 2001). No entanto, como discutido neste trabalho e por outros autores, a análise dos dados pode ser mais complexa dada a possibilidade de inúmeras formas e padrões de ausência de informação. As técnicas estatísticas disponíveis podem ser extremamente úteis nessa tarefa, porém o uso apropriado depende do entendimento do contexto e das possíveis causas da perda de informação, do uso de testes estatísticos para avaliação do tipo de mecanismo de ausência de dados e da identificação de variáveis relacionadas com a ausência e que podem ser utilizadas na redução de vieses nas estimativas.

REFERÊNCIAS

Alabadla, M., Sidi, F., Ishak, I., Ibrahim, H., Affendey, L. S., Ani, Z. C., Jabar, M. A., Bukar, U. A., Devaraj, N. K., Muda, A. S., Tharek, A., Omar, N., & Jaya, M. I. M. (2022). Systematic review of using machine learning in imputing missing values. *IEEE Access*, *10*, 44483-44502. https://doi.org/10.1109/ACCESS.2022.3160841

Allison, P. D. (2001). *Missing data*. Sage.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analysis. *Journal of School Psychology, 48*(1), 5-37. https://doi.org/10.1016/j.jsp.2009.10.001

Bello, L., & Britto, V. (2024, 22 março). Uma em cada quatro mulheres de 15 a 29 anos não estudava e nem estava ocupada em 2023. *Agência IBGE Notícias*. https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/39531-uma-em-cada-quatro-mulheres-de-15-a-29-anos-nao-estudava-e-nem-estava-ocupada-em-2023

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487-508. https://doi.org/10.3102/0034654314532697

Collins, L. M., Schafer, J. L., & Kam, C-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330-351. https://doi.org/10.1037//1082-989X.6.4.330

Davis, R., Occhipinti, S., & Jones, L. (2018). Managing missing data: Concepts, theories, and methods. In P. Brough (Ed.), *Advanced research methods for applied psychology* (pp. 187-200). Routledge.

Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable dropout. *Statistics in Medicine*, *22*(16), 2553-2575. https://doi.org/10.1002/sim.1475

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods, 16*(1), 1-16. https://doi.org/10.1037/a0022640

Enders, C. K. (2022). *Applied missing data analysis*. Guilford.

Enders, C. K. (2023). Missing data: An update on the state of the art. *Psychological Methods*, *30*(2), 322-339. https://doi.org/10.1037/met0000563

Ferrão, M. E., & Prata, P. (2019). Computing topics on multiple imputation in Big Identifiable Data using R: An application to educational research. In *19th International Conference on Computational Science and Its Applications: ICCSA 2019* (Part. 3, pp. 12-24). Springer.

Ferrão, M. E., Prata, P., & Alves, M. T. G. (2020). Multiple imputation in big identifiable data for educational research: An example from the Brazilian education assessment system. *Ensaio: Avaliação e Políticas Públicas em Educação, 28*(108), 599-641. https://doi.org/10.1590/S0104-40362020002802346

Ferreira, M. E. (2022). Evasão escolar no ensino médio: Possíveis causas e soluções. *RCMOS – Revista Científica Multidisciplinar O Saber, 1*(1), 310-315. https://doi.org/10.51473/rcmos.v2i1.277

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis*. Chapman & Hall. https://doi.org/10.1201/9781420011579

Franco, C. (2001). O SAEB – Sistema de Avaliação da Educação Básica: Potencialidades, problemas e desafios. *Revista Brasileira de Educação,* (17), 127-133. https://www.scielo.br/j/rbedu/a/qCYrZ7vVQYFH7fRXBhBZ5Nm/abstract/?lang=pt

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549-576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206-213. https://doi.org/10.1007/s11121-007-0070-9

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*(1), 64-78. https://doi.org/10.1037/1082-989X.2.1.64

Ismail, A. R., Abidin, N. Z., & Maen, M. K. (2022). Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. *Journal of Robotics and Control (JRC), 3*(2), 143-152. https://doi.org/10.18196/jrc.v3i2.13133

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychobiology, 45*(4), 1195-1199. https://doi.org/10.1037/a0015665

Leon, F. L. L. de, & Menezes, N. A., Filho (2002). Reprovação, avanço e evasão escolar no Brasil. *Pesquisa e Planejamento Econômico, 32*(3), 417-452. http://repositorio.ipea.gov.br/handle/11058/4286

Little, R. J. (2024). Missing data analysis. *Annual Review of Clinical Psychology, 20,* 149-173. https://doi.org/10.1146/annurev-clinpsy-080822-051727

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford.

Ministério da Educação (MEC). (2024, 22 fevereiro). Ensino médio tem maior taxa de evasão da educação básica. *agência gov.* https://agenciagov.ebc.com.br/noticias/202402/ensino-medio-tem-maior-taxa-de-evasao-da-educacao-basica

Neri, M. (2009). *Motivos da evasão escolar.* Fundação Getulio Vargas.

Occhipinti, S. (2024). Missing data. In P. Brough (Ed.), *Advanced research methods for applied psychology* (pp. 211-223). Routledge.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation, 7*(4), 353-383. https://doi.org/10.1076/edre.7.4.353.8937

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd edition). Sage.

Rousseau, M., Simon, M., Bertrand, R., & Hachey, K. (2012). Reporting missing data: A study of selected articles published from 2003-2007. *Quality & Quantity, 46*(5), 1393-1406. https://doi.org/10.1007/s11135-011-9452-y

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. https://doi.org/10.2307/2335739

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

Secretaria da Educação. (2011). *SPAECE 2011: Boletim Pedagógico Matemática – Ensino Médio* (Vol. 3). CAEd; Governo do Estado do Ceará. https://prototipos.caeddigital.net/arquivos/ce/colecoes/2011/BOLETIM_SPAECE_VOL%203_MT_3%20EM.pdf

Seu, K., Kang, M.-S., & Lee, H. (2022). An intelligent missing data imputation techniques: A review. *JOIV: International Journal on Informatics Visualization*, *6*(1-2), 278-283. http://dx.doi.org/10.30630/joiv.6.1-2.935

Shirasu, M. R. (2014). *Determinantes da evasão e repetência escolar no Ceará* [Dissertação de mestrado, Universidade Federal do Ceará]. Repositório Institucional UFC. http://repositorio.ufc.br/handle/riufc/15223

Silva, J. L. P. (2013). *Métodos de imputação múltipla para GEE em estudos longitudinais* [Dissertação de mestrado, Universidade Federal de Minas Gerais]. Repositório Institucional da UFMG. https://hdl.handle.net/1843/BUOS-8GHJRP

Simões, A. (2014). *Acesso e evasão na educação básica: As perspectivas da população de baixa renda no Brasil*. Ministério do Desenvolvimento Social e Assistência Social, Família e Combate à Fome.

Soares, T. M., Fernandes, N. da S., Nóbrega, M. C., & Nicolella, A. C. (2015). Fatores associados ao abandono escolar no ensino médio público de Minas Gerais. *Educação e Pesquisa*, *41*(3), 757-772. https://doi.org/10.1590/S1517-9702201507138589

Vinha, L. G. do A. (2016). *Estudos longitudinais e tratamento de dados ausentes em avaliações educacionais* [Tese de doutorado, Universidade de Brasília]. Repositório Institucional UnB. http://repositorio.unb.br/handle/10482/20204

Vinha, L. G. do A., & Laros, J. A. (2018). Dados ausentes em avaliações educacionais: Comparação de métodos de tratamento. *Estudos em Avaliação Educacional*, *29*(70), 156-187. https://doi.org/10.18222/eae.v0ix.3916

Wærsted, M., Børnick, T. S., Twisk, J. W. R., & Veiersted, K. B. (2018). Simple descriptive missing data indicators in longitudinal studies with attrition, intermittent missing data and a high number of follow-ups. *BMC Research Notes*, *11*, 1-7. https://doi.org/10.1186/s13104-018-3228-6

NOTA: As contribuições de cada autor para o desenvolvimento do artigo foram as seguintes: Luís Gustavo do Amaral Vinha – conceitualização; curadoria e análise de dados; redação do manuscrito original, revisão e aprovação da versão final do trabalho. Jacob Arie Laros – conceitualização; curadoria e validação de dados; redação: revisão e aprovação da versão final do trabalho.