**Construção da capacidade avaliativa, uso da avaliação e desempenho de escolas: uma análise com dados do SAEB**


**Resumo**

O objetivo desta pesquisa é analisar como variáveis relacionadas ao uso da avaliação por diretores de escolas e por redes municipais de educação influenciam no desempenho escolar medidos em testes de larga escala de Língua Portuguesa (LP) e Matemática (MAT) no 5º ano do Ensino Fundamental. Os dados foram obtidos no Serviço de Acesso a Dados Protegidos do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, órgão vinculado ao Ministério da Educação. Foram utilizados modelos lineares multinível para análise de mais de 33 mil escolas em 4500 municípios brasileiros. Os resultados indicam que o uso da avaliação pelas escolas e pelas redes municipais de educação têm um efeito positivo e significativo no desempenho em LP e MAT, apesar do tamanho de efeito ser relativamente reduzido. As redes municipais de educação que fazem maior uso da avaliação possuem maior desempenho escolar em LP e MAT, indicando que esforços de construção da capacidade avaliativa devem ser concentrados mais nelas do que nas escolas que as compõem.



**Palavras-chave:** capacidade avaliativa, uso da avaliação, desempenho escolar, SAEB.


## 1. Introdução

O Brasil possui sistemas avaliativos ligados à educação implantados há alguns anos, como o Sistema de Avaliação da Educação Básica (SAEB), Sistema de Avaliação do Ensino Superior (Sinaes) e Sistema de Avaliação da Pós-graduação (conhecido informalmente como avaliação Capes), sendo significativo o montante de recursos investidos neles. Cada sistema tem sua lógica, seu formato, sua política avaliativa.

O esforço continuado para desenvolver, conduzir e sustentar os processos organizacionais associados aos sistemas avaliativos, incluindo suas rotinas burocráticas e operacionais é definido como a construção da capacidade avaliativa (ECB, do inglês, *Evaluation Capacity Building*) e visa garantir a qualidade das avaliações (Stockdill; Baizerman; Compton, 2002). O conceito de ECB, que historicamente concentrou-se nas competências individuais e organizacionais para realizar avaliações de alta qualidade (Preskill; Boyle, 2008),

foi expandido para incluir também a capacidade para usar as avaliações, ou seja, a habilidade para integrar a avaliação nos seus processos de tomada de decisão (Bourgeois *et al.*, 2016; Bourgeois; Cousins, 2009). Assim, para além de ter sistemas avaliativos institucionalizados (Leeuw; Furubo, 2008), é fundamental que eles forneçam informações que possam ser usadas por tomadores de decisão (sejam professores, diretores, secretários, políticos etc.) e que culminem na melhoria dos indicadores finais (Neuman *et al.*, 2013), no caso, aqueles ligados à educação.

O conceito de uso da avaliação (EU, do inglês, *Evaluation Use*) foi inicialmente atrelado às ações imediatas e diretas derivadas dos resultados avaliativos. Este tipo de uso, denominado instrumental, é tido como o mais observável e esperado no curto prazo (Acree; Chouinard, 2020). Contudo, o conceito de EU foi expandido para abarcar usos associados com as alterações de compreensão sobre o objeto avaliado e sobre a organização a qual ele pertence. Ou seja, a incorporação de novos entendimentos advindos da avaliação é um tipo de uso, denominado conceitual, e prescinde de qualquer ação ou decisão direta ou imediata. Há ainda o uso político e persuasivo da avaliação, chamado de uso simbólico (Alkin; King, 2016).

Em suma, o uso da avaliação faz parte da construção da capacidade avaliativa de uma organização ou sistema, assim como a capacidade da mesma de conduzir a avaliação. A ampliação dos usos seria, então, um indicador associado à ECB e, em última instância, de institucionalização da política avaliativa (EP, do inglês, *Evaluation Policy*) (Al Hudib; Cousins, 2022) o que, por fim, atenderia ao princípio maior de melhoria da sociedade (Kirkhart, 2000).

A relação entre a EP, a ECB e o EU pode ser explicada por diferentes elementos em níveis distintos. O nível macro refere-se ao contexto (político, social, cultural e econômico) e a política avaliativa. O nível meso abarca a organização, ou sistema de organizações, com sua capacidade de conduzir e usar avaliações. E no nível micro tem-se os indivíduos (stakeholders) com seus papéis e interrelações (Al Hudib; Cousins, 2022). As organizações, nível intermediário de análise, podem construir sua capacidade de executar e usar avaliações de qualidade a partir dos indivíduos. Essa capacidade pode ser restringida ou ampliada em função da EP e do contexto no qual a organização está inserida.

A reflexão a partir de níveis distintos é importante para a compreensão do posicionamento das diversas variáveis no campo empírico. Assim, esta pesquisa tem como objeto de análise o SAEB, implantado na década de 1990. Além das provas em larga escala que ajudam a compor as notas das escolas, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), responsável pela execução do SAEB, coleta uma série de dados

relevantes advindos de professores, diretores de escola e secretários municipais de educação. Alguns podem ser associadas como *proxies* para a compreensão do EU, e consequentemente da ECB, no âmbito da educação básica brasileira.

## 2. Objetivos geral e específicos

Tendo como foco analítico os níveis meso (organizacional) e micro (individual) e considerando que a capacidade de conduzir avaliações está institucionalizada, esta pesquisa objetiva analisar como variáveis relacionadas ao EU nos níveis meso (rede municipal de educação) e micro (diretores de escolas) influenciam no desempenho escolar medidos em Língua Portuguesa (LP) e Matemática (MAT) no 5º ano do Ensino Fundamental I (EF-1). A operacionalização das variáveis é apresentada na seção de métodos.

Os objetivos específicos são:

Analisar o quanto de variabilidade no desempenho das escolas em LP e MAT (5º ano) advém da rede municipal de educação na qual ela está inserida;

Analisar a influência dos usos da avaliação por parte de diretores de escolas no desempenho das escolas em LP e MAT (5º ano);

Analisar a influência dos usos da avaliação por parte de redes municipais de educação no desempenho das escolas em LP e MAT (5º ano).

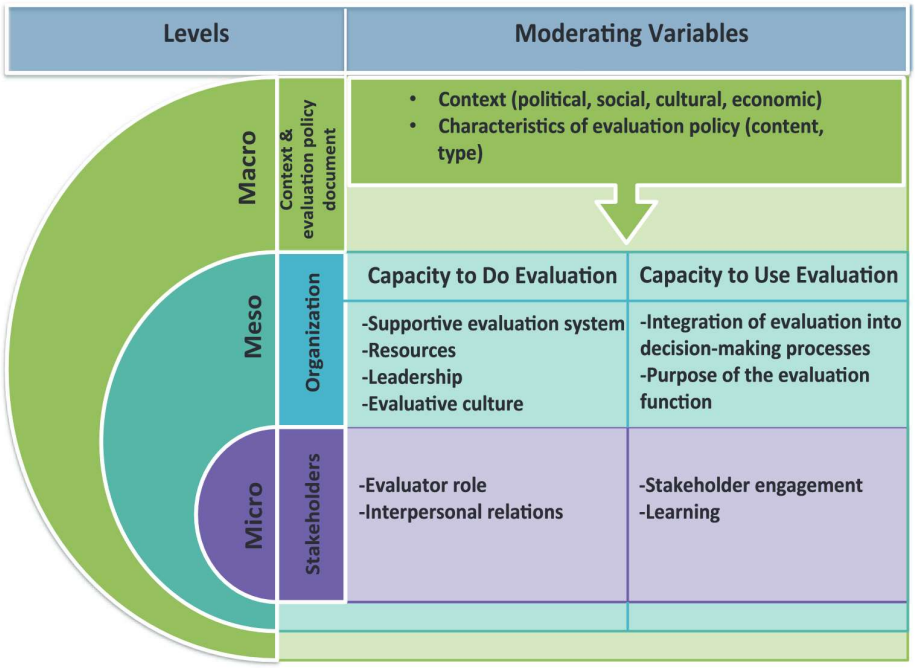
Convém ressaltar que as pesquisas conduzidas no Brasil acerca do SAEB, agrupadas em fatores extra e intraclasse, eficácia escolar, desigualdade educacional, políticas públicas educacionais, qualidade e equidade na educação básica, indicadores educacionais, mineração de dados educacionais e *accountability* educacional (Gusmão; Amorim, 2023), não utilizaram os dados de interesse dessa pesquisa, o que evidencia seu ineditismo. Ademais, entende-se que a verificação empírica da associação de variáveis, ainda que aproximadas, de EU e, consequentemente de ECB, com desempenho escolar é uma contribuição importante à literatura de avaliação, incluindo EP, ECB e EU, carente de pesquisas (Dillman; Christie, 2017), notadamente estudos quantitativos e com grandes bancos de dados (Daigneault, 2014). Por fim, entende-se que essa pesquisa pode trazer uma contribuição prática com possível influência nas definições da EP educacional.

## 3. Referencial teórico

As diretrizes e princípios que guiam as ações e decisões das organizações e das pessoas envolvidas em processos avaliativos fazem parte da política avaliativa, que pode ser escrita/explícita ou não escrita/implícita (Al Hudib; Cousins, 2022; Trochim, 2009). Os estudos sobre EP ainda estão em fase de desenvolvimento teórico (Dillman; Christie, 2017), notadamente em relação a seus efetivos impactos no desenvolvimento de uma cultura avaliativa organizacional e na melhoria do que está sendo avaliado (Mark; Cooksy; Trochim, 2009). A EP é um importante mecanismo de comunicação da avaliação, envolvendo como a avaliação deve ser conduzida, quais recursos são necessários, quem serão os responsáveis e demais detalhes (Trochim, 2009). Logo, a EP é um elemento central para o desenvolvimento da capacidade avaliativa (Mark; Cooksy; Trochim, 2009).

Essa capacidade pode ser apresentada a partir de três níveis (ver figura 1): o contexto macro, o organizacional (meso) e o individual (micro). No nível macro, os contextos político, social, cultural e econômico, além das próprias características da política avaliativa, influenciam a capacidade das organizações de conduzirem e usarem avaliações (ECB). No nível organizacional (meso), a capacidade para conduzir avaliações de alto nível é influenciada pelo volume de recursos, liderança, cultura avaliativa e suporte do sistema avaliativo. Já a capacidade de usar as avaliações é influenciada pelo seu propósito e pela sua integração ao processo de tomada de decisões. E no nível micro, a capacidade de conduzir avaliações é influenciada pelo papel do avaliador e pelas relações interpessoais que ocorrem no âmbito da avaliação, enquanto que a capacidade de usar as avaliações é influenciada pelo engajamento dos stakeholders e pelo processo de aprendizado que ocorre quando da execução das mesmas (Al Hudib; Cousins, 2022).

**Figura 1: Modelo de análise para ECB.**



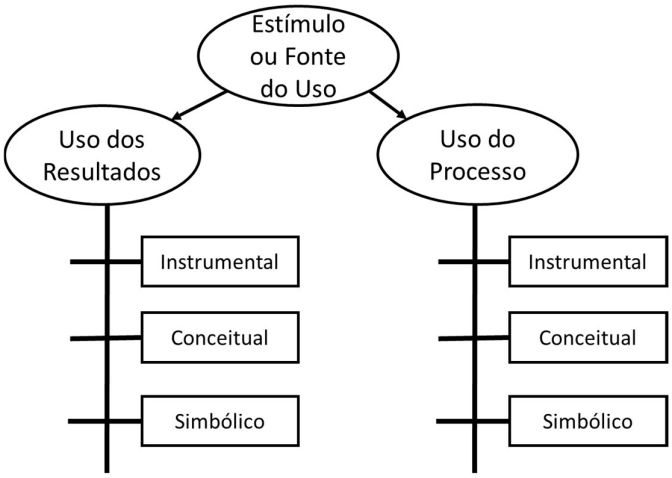

Fonte: Al Hudib e Cousins (2022).

No nível organizacional, interessa para essa pesquisa, especificamente, a capacidade de usar a avaliação. A ECB não é completa se não há uso efetivo da avaliação (Al Hudib; Cousins, 2022). Não usar significaria, no fim das contas, desperdício de recursos (Neuman *et al.*, 2013). A capacidade de usar avaliações envolve o conhecimento dos indivíduos sobre avaliação (letramento em avaliação) e a integração da avaliação com o processo decisório no nível organizacional (Bourgeois; Cousins, 2013). Ainda que haja uso tácito e não deliberado da avaliação, uma sistemática exemplar de ECB deve compor o uso como parte integrante da política avaliativa.

Os usos da avaliação podem advir de duas fontes: a participação no processo avaliativo (*process use*) e os resultados da avaliação (*findings use*). E três tipos de usos são destacados na literatura: uso instrumental, uso conceitual e uso simbólico (ver Figura 2). O uso instrumental é aquele uso imediato, direto, visível em decisões e ações. O uso conceitual refere-se a como a avaliação afetou a forma de pensar (da própria avaliação ou do objeto avaliado) de uma ou mais pessoas, sem necessariamente envolver ações diretas. E o uso simbólico está associado ao uso político, persuasivo ou legitimador da avaliação (Alkin; King, 2017; Alkin; Taut, 2002). Assim, a título de exemplo, tem-se uso instrumental advindo dos resultados quando um indivíduo decide modificar um determinado processo na organização após ter tido acesso aos resultados

de uma avaliação. Outro exemplo é quando um indivíduo, por ter participado do processo avaliativo, modifica seu entendimento sobre o funcionamento do objeto avaliado, sem necessariamente tomar nenhuma ação, o que seria uso conceitual advindo do processo.

**Figura 2: Fontes e tipos de usos da avaliação.**




Fonte: Alkin e Taut (2002).

Importante destacar que o uso, independente do seu tipo ou fonte, é capaz de gerar conhecimento e, em última instância, melhorar o programa avaliado (Henry; Mark, 2003; Kirkhart, 2000; Mark; Henry, 2004). Além disso, conhecer os usos ajuda na identificação de usos não intencionais (Van Thiel; Leeuw, 2002) ou usos indevidos (Alkin; King, 2017), muitas vezes associados a erros de interpretação por falta de letramento avaliativo ou comportamento antiético. Dessa forma, entende-se que é fundamental compreender os usos que são feitos de uma avaliação para construir capacidade avaliativa (ECB), garantir as definições concebidas na política avaliativa e, em última instância, promover a melhoria no sistema/programa avaliado.

## 4. Procedimentos metodológicos

Essa pesquisa é quantitativa e utiliza modelos de regressões lineares multinível. Os dados da pesquisa foram obtidos junto ao Serviço de Acesso a Dados Protegidos (Sedap) do INEP. Como os dados permitem a identificação de pessoas, o acesso e a extração precisam de autorização do Sedap, o que foi obtido em outubro de 2024 (processo 23036.008878/2024-67). Foram utilizados os dados do SAEB 2021, último disponível à data.

Com base no objetivo proposto, tem-se como variável dependente o desempenho da escola. Operacionalmente, são duas variáveis dependentes, que são analisadas em modelos de regressão distintos, mas similares. São elas: média em LP do 5º ano e média em MAT do 5º ano, ambas agregadas no nível da escola. O INEP utiliza a Teoria da Resposta ao Item (TRI), sendo as notas analisadas a partir de escalas de proficiência. Para LP, a escala do 5º ano vai do nível zero (notas entre 0 a 125) ao nível 9 (notas acima de 325) com aumentos de 25 pontos em cada nível. Para MAT a escala do 5º ano vai do nível zero (notas entre 0 e 125) ao nível 10 (notas acima de 350), também com aumentos de 25 pontos por nível (Brasil, 2020).

As variáveis independentes de interesse são aquelas que estão associadas com o conceito de EU e que foram obtidas pelo INEP nos questionários aplicados junto aos diretores de escolas e aos secretários municipais de educação (ver Quadro 1). Tais variáveis não são escalas validadas como outros instrumentos apresentados na literatura (Bourgeois *et al.*, 2016; Paixão; Rodriguez, 2023), mas podem servir como *proxies* simplificadas.

Foram considerados indicativos de EU no nível da escola (micro): a) a consideração, no projeto político-pedagógico (PPP) da escola, dos resultados previamente obtidos em avaliações externas; e b) a existência de metas de alcance de indicadores externos (como o Índice de Desenvolvimento da Educação Básica, IDEB, gerado pelo SAEB). Especificamente, entende-se que a construção de um PPP a partir dos resultados de avaliações prévias e a inclusão de metas associadas a indicadores de desempenho obtidos em avaliações em larga escala constituem usos instrumentais dos resultados da avaliação (Alkin; Taut, 2002). Como são dois itens dicotômicos, a variável foi agregada pela soma dos mesmos (mínimo possível 0, máximo 2). A variável foi denominada de uso da avaliação pela escola (EU-Escola).

Foram considerados *proxies* de EU no nível organizacional (meso) a utilização dos resultados do IDEB, seja para produção de materiais ou identificação de escolas carentes de auxílio (ver Quadro 1). Como são sete itens dicotômicos, a variável também foi agrupada pela soma (mínimo possível 0, máximo 7). Os itens constam no questionário respondido pelo(a) secretário(a) municipal de educação e foram considerados como usos instrumentais dos resultados da avaliação no nível organizacional. Essa variável foi denominada uso da avaliação pela rede educacional (EU-Rede). Como o uso é um dos componentes da ECB, e considerando o nível organizacional, pode-se associar um maior/menor uso com o maior/menor desenvolvimento da capacidade avaliativa (Al Hudib; Cousins, 2022).

**Quadro 1: Resumo das variáveis independentes de interesse.**

| Nível | Variável | Forma de mensuração |
|---|---|---|
| Escola / Diretor(a) | - PPP da escola considera os resultados de avaliações externas (SAEB, estaduais, municipais etc.).<br>- Existência de metas de alcance de indicadores externos (Ideb, índices estaduais ou municipais). | Binárias (Sim; Não)<br><br>Agregadas pela soma (mínimo 0, máximo 2) |
| Rede / Secretaria municipal de educação | Formas de utilização dos resultados do IDEB:<br>- Subsídios para a formação continuada de professores.<br>- Subsídios para a avaliação de programas ou projetos da Secretaria de Educação.<br>- Produção de materiais didáticos e pedagógicos.<br>- Recompensa para escolas com melhores resultados.<br>- Auxílio para unidades escolares com resultados inferiores.<br>- Subsídios para pagamento de bonificação para professores.<br>- Criação de critérios para remanejamento de diretores. | Binárias (Sim; Não)<br><br>Agregadas pela soma (mínimo 0, máximo 7) |

Fonte: Elaborado pelo autor.

Pesquisas indicam um conjunto relativamente amplo de variáveis que influenciam no desempenho de alunos e escolas (Brooke; Soares, 2008; Fryer, 2012; Scheerens, 2016). A incorporação de outras covariáveis ao modelo visa captar o efeito das variáveis de interesse após "controle". Contudo, a inclusão de covariáveis deve estar respaldada na racionalidade teórica e empírica (Bernerth; Aguinis, 2016; Spector; Brannick, 2011), além da interpretabilidade dos resultados (Becker, 2005). Os dados do SAEB permitem a adição de muitas variáveis de controle, contudo isso implicaria no aumento da complexidade dos modelos. Assim, optou-se por incluir apenas a variável de nível socioeconômico, dado que a literatura a apresenta como uma das mais importantes na predição do desempenho (Brooke; Soares, 2008; Scheerens, 2016). O INEP calcula um Indicador de Nível Socioeconômico (INSE) a partir de 17 itens respondidos pelos estudantes. Os itens envolvem a escolaridade dos pais e a existência de determinados equipamentos na casa do estudante, como, por exemplo, geladeira, computador, televisão, banheiro, carro, celular, internet e quantidade de quartos na casa. O indicador final possui oito níveis, do menor ao maior nível socioeconômico e são agregados por escola. O INEP conduziu estudo comparando o INSE com outras medidas de nível socioeconômico encontrando correlações elevadas e significativas (Brasil, 2023), o que indica sua validade.

Os procedimentos de análise envolveram a elaboração de modelos de regressão multinível para cada variável dependente. A modelagem linear multinível (MLM, do inglês *Multilevel Linear Modeling*) é considerada a técnica adequada porque as escolas estão

agrupadas em redes/municípios (*clusters*). Assim, considera-se as informações coletadas da escola/diretor(a) como sendo o primeiro nível e da rede/secretaria como o segundo. Importante destacar que a MLM possui a vantagem de considerar, no modelo, a estrutura hierárquica dos dados e, consequentemente, gerar parâmetros não viesados. Ou seja, quando a estrutura multinível não é levada em consideração, problemas estatísticos e conceituais emergem. Primeiro, pode-se incorrer no risco de analisar dados em um nível e tirar conclusões para outro nível. Segundo, os erros padrão podem ser subestimados e elevar a chance de erro do tipo I (encontrar associações significativas quando elas não existem). Além disso, escolas de uma mesma rede tendem a possuir características similares e, portanto, não podem ser consideradas como independentes (Chen; Chen, 2021; Raudenbush; Bryk, 2002).

Para cada variável dependente foram gerados quatro modelos. O primeiro, denominado *unconditional model*, serviu de base para a estimação da correlação intraclasse (ICC) e verificação da adequação da MLM. No modelo 2 foi adicionada a variável referente ao uso da avaliação por parte do Diretor da escola (de nível 1). No modelo 3 adicionou-se a variável referente ao uso da avaliação por parte do Secretário Municipal de Educação (nível 2). No modelo 4 foi adicionada a variável de controle referente ao nível socioeconômico agregado da escola. A adição das variáveis em modelos distintos permite a análise dos coeficientes e tamanhos de efeito (Nakagawa; Johnson; Schielzeth, 2017) em cada etapa, sendo a geração de modelos comparativos uma prática da MLM (Chen; Chen, 2021).

As variáveis do nível 1, EU-Escola e INSE, foram centralizadas pela média do grupo (rede municipal de educação) a que pertencem (*group-mean centering*). A variável de nível 2, EU-Rede, foi centralizada pela média geral (*grand-mean centering*). A centralização de variáveis é um mecanismo muito utilizado nas modelagens multinível e visam facilitar a interpretação dos resultados (Luo *et al.*, 2021; Yaremych; Preacher; Hedeker, 2023).

Os modelos podem ser comparados de forma descritiva a partir dos indicadores *Akaike Information Criterion* (AIC), *Bayesian Information Criterion* (BIC) e *Deviance*, sendo que um menor valor indica melhor ajustamento do modelo (Burnham; Anderson, 2004). Para cada modelo final (modelo 4), os pressupostos de heterocedasticidade e normalidade dos resíduos foram graficamente verificados (Chen; Chen, 2021; Khine, 2022). As análises foram conduzidas utilizando o software R com os pacotes lme4 (Bates *et al.*, 2015), lmerTest (Kuznetsova; Brockhoff; Christensen, 2017), performance (Lüdecke *et al.*, 2021) e sjPlot (Lüdecke, 2024).

O SAEB 2021 foi censitário para as escolas públicas com mais de 10 matrículas. A população de referência do INEP, a qual exclui escolas com menos de 10 matrículas, escolas exclusivas de turmas multisseriadas e escolas indígenas, era de 46.004 escolas públicas do EF1 (Brasil, 2021). Foram mantidas na base apenas as escolas públicas e as que possuem PPP. A base final para o EF1 foi de 33.388 escolas em 4.500 municípios.

## 5. Resultados

A Tabela 1 apresenta as estatísticas descritivas para as variáveis utilizadas na análise. As notas de LP variaram entre zero e 319,25, enquanto em MAT o máximo foi de 338,78. Para EU-Escola a mediana ficou no valor máximo, o que indica que a grande maioria dos diretores das escolas informa fazer uso da avaliação. Para a rede educacional, a média foi de 2,74 e o coeficiente de variação foi de 0,562, evidenciando uma alta dispersão relativa em torno da média.

**Tabela 1: Estatísticas descritivas.**

| Variável | Min. | Mediana | Média | Máx. | Desvio padrão | Coef. de Variação |
|---|---|---|---|---|---|---|
| Nota LP | 0 | 200,17 | 198,03 | 319,25 | 29,47 | 0,149 |
| Nota MAT | 0 | 207,31 | 207,19 | 338,78 | 29,73 | 0,143 |
| EU - Escola | 0 | 2,00 | 1,90 | 2,00 | 0,37 | 0,197 |
| EU - Rede | 0 | 3,00 | 2,74 | 7,00 | 1,54 | 0,562 |
| INSE | 0 | 3,16 | 3,00 | 6,00 | 1,08 | 0,360 |

Em seguida, obteve-se o valor da ICC no modelo 1. Para LP, a ICC foi de 0,578, enquanto que para MAT foi igual a 0,610. Significa que 57,8% e 61% da variação no desempenho das escolas em LP e MAT, respectivamente, advém das redes que as compõem. Ou seja, as variações entre as redes municipais de educação são muito mais relevantes do que as variações entre as escolas de uma mesma rede.

Na Tabela 2 são apresentadas as estimativas obtidas em cada um dos modelos para LP. O valor do intercepto apresentado no modelo 1 indica que a nota média em LP entre todas as escolas foi de 196,58. No modelo 2, adicionou-se o uso da avaliação pela escola, sendo o coeficiente positivo e significativo. O aumento do uso da avaliação por uma escola de uma determinada rede municipal de educação em um ponto implica no aumento da nota em LP em 2,59 pontos a mais que a média da sua rede. No modelo 3, o aumento no uso da avaliação pela rede educacional em um ponto resulta no aumento da nota em LP em 3,54 pontos a mais que a média geral das redes, não havendo modificação no coeficiente de EU-Escola. Já no modelo 4,

a adição do INSE altera os parâmetros de uso da avaliação, seja da escola ou da rede. O aumento do INSE em um ponto (nível) resulta em um aumento da nota em LP em 13,6 pontos acima da média da sua rede. Os coeficientes para EU-Escola e EU-Rede foram reduzidos para 0,58 e 2,78, porém ambos se mantiveram positivos e significativos.

**Tabela 2: Estatísticas para os modelos de Língua Portuguesa.**

| Nível | Parâmetro | Modelos: Estimativas (erros padrão) | | | |
|---|---|---|---|---|---|
| | | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 |
| Escola | Intercepto | 196,58*** (0,406) | 196,58*** (0,406) | 196,86*** (0,410) | 198,70*** (0,337) |
| | EU-Escola | | 2,59*** (0,332) | 2,59*** (0,333) | 0,58*** (0,248) |
| | INSE | | | | 13,6*** (0,159) |
| Rede de Educação Municipal | EU-Rede | | | 3,54*** (0,819) | 2,78*** (0,673) |
| Efeitos aleatórios | Rede | 587,54 | 587,89 | 584,43 | 408,90 |
| | Resíduo | 429,41 | 428,50 | 428,54 | 233,30 |
| Variância explicada | $R^2$ Marginal | 0,000 | 0,001 | 0,004 | 0,077 |
| | $R^2$ Condicional | 0,578 | 0,579 | 0,578 | 0,665 |
| Comparação de modelos | # Parâmetros | 3 | 4 | 5 | 6 |
| | AIC | 305.598 | 305.539 | 305.522 | 248.398 |
| | BIC | 305.623 | 305.573 | 305.565 | 284.449 |
| | *Deviance* | 305.592 | 305.531 | 305.512 | 284.386 |
| N | Escolas | 33.388 | 33.388 | 33.388 | 33.186 |
| | Redes | 4.500 | 4.500 | 4.500 | 4.466 |

Nota. * $p < 0.05$, *** $p < 0.001$.

Os valores obtidos para efeitos aleatórios indicam que a adição de variáveis aos modelos fez com que a variabilidade não explicada pelas mesmas fosse reduzida. Contudo, os tamanhos dos efeitos obtidos foram baixos, como 0,1% (modelo 2), 0,4% (modelo 3) e 7,7% (modelo 4). Esses resultados indicam que as variáveis de uso da avaliação, seja no nível da escola ou no nível da rede municipal de educação, explicam pouco do resultado em LP (0,4% no modelo 3).

No modelo 4 o tamanho do efeito foi maior, indicando que o INSE é bastante relevante para um aumento em LP, mais do que EU-Escola e EU-Rede. Os valores apresentados para AIC, BIC e *Deviance* foram reduzidos no modelo 4, indicando este como o de melhor ajuste.

Na Tabela 3 são apresentadas as estimativas para os modelos com base na nota de Matemática. O intercepto apresentado no modelo 1 indica que a média de MAT entre todas as escolas foi de aproximadamente 207 pontos. As adições de EU-Escola no modelo 2 e EU-Rede no modelo 3 indicam que o aumento do uso da avaliação por uma escola de uma determinada rede municipal de educação em um ponto resulta em um aumento na nota em MAT em 2,52 pontos acima da média da sua rede, enquanto que o aumento em um ponto no uso da avaliação da rede impactaria no aumento em 4,15 pontos no desempenho em MAT acima da média geral. Ao adicionar o INSE os coeficientes de EU-Escola e EU-Rede ainda se mantêm positivos e significativos (0,77 e 3,38, respectivamente). O aumento em um nível no INSE de uma escola em uma determinada rede teria um impacto na elevação da nota em MAT em 12 pontos acima da média daquela rede. Assim como em LP, o INSE mostra-se uma variável de maior impacto do que ambas as variáveis de uso da avaliação.

**Tabela 3: Estatísticas para os modelos de Matemática.**

| Nível | Parâmetro | Modelos: Estimativas (erros padrão) | | | |
|---|---|---|---|---|---|
| | | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 |
| Escola | Intercepto | 207,02*** (0,421) | 207,03*** (0,421) | 207,35*** (0,426) | 209,30*** (0,345) |
| | EU-Escola | | 2,51*** (0,326) | 2,52*** (0,326) | 0,77** (0,241) |
| | INSE | | | | 12,00*** (0,154) |
| Rede de Educação Municipal | EU-Rede | | | 4,15*** (0,849) | 3,38*** (0,688) |
| Efeitos aleatórios | Rede | 648,02 | 648,38 | 643,70 | 435,50 |
| | Resíduo | 413,66 | 412,81 | 412,84 | 220,20 |
| Variância explicada | $R^2$ Marginal | 0,000 | 0,001 | 0,004 | 0,061 |
| | $R^2$ Condicional | 0,610 | 0,611 | 0,609 | 0,685 |
| Comparação de modelos | # Parâmetros | 3 | 4 | 5 | 6 |
| | AIC | 304.839 | 304.781 | 304.759 | 282.936 |

| | | | | | |
|---|---|---|---|---|---|
| | BIC | 304.864 | 304.815 | 304.802 | 282.986 |
| | *Deviance* | 304.833 | 304.773 | 304.749 | 282.924 |
| N | Escolas | 33.388 | 33.388 | 33.388 | 33.186 |
| | Redes | 4.500 | 4.500 | 4.500 | 4.466 |

Nota. * $p < 0.05$, *** $p < 0.001$.

No geral, os resultados de MAT espelham os de LP. Os valores de variância dos efeitos aleatórios também foram reduzidos no modelo 4. O tamanho do efeito obtido no modelo 2 foi de apenas 0,01% e no modelo 3 de 0,04%. Isso significa que as variáveis EU-Escola e EU-Rede sozinhas explicam até 0,4% da variação no desempenho das escolas em LP. No modelo 4 o tamanho do efeito foi de 6,1%, o que novamente indica a importância do INSE. Os indicadores de comparação dos modelos também foram reduzidos, notadamente no modelo 4, indicando este como o de melhor ajustamento.

## 6. Discussão

As pesquisas sobre uso da avaliação têm avançado na identificação dos tipos de usos, fontes e fatores que afetam o uso (Alkin; King, 2016). Contudo, nem sempre é simples associar um maior uso da avaliação com o seu respectivo efeito em termos de desempenho do objeto avaliado (Kirkhart, 2000). Essa pesquisa buscou associar os usos da avaliação por escolas (EU-Escola) e redes municipais de educação (EU-Rede) com o desempenho médio obtido pelas escolas em língua portuguesa e matemática no 5º ano do Ensino Fundamental.

Os resultados da correlação intraclasse (ICC) obtidos nos modelos 1 para LP (0,578) e MAT (0,610) no 5º ano indicam que a variabilidade no desempenho das escolas advém, em sua grande maioria, das redes municipais de educação que as compõem. Ou seja, as maiores variações dão-se entre redes, não entre escolas pertencentes a uma mesma rede. Considerando os valores de ICC, é necessário explorar mais as diferenças de usos da avaliação entre as redes municipais de educação para melhor compreender como elas afetam o desempenho das escolas.

Os usos feitos pelos diretores e secretários municipais de educação estão positivamente associados com o desempenho em LP e MAT. Um diretor de escola que usar mais a avaliação vai conseguir melhorar o desempenho da escola em LP e MAT (5º ano), mas não tanto. Um secretário municipal de educação que usar mais a avaliação vai conseguir melhorar o desempenho das suas escolas da sua rede em LP e MAT (5º ano) comparativamente mais do

que o diretor. Ambas as variáveis explicam aproximadamente 0,4% (modelos 3) do desempenho em LP e MAT (5º ano), o que pode ser considerado um tamanho do efeito reduzido, quando analisado em relação ao indicador de nível socioeconômico. Adicionar o INSE nos modelos 4 fez com que o tamanho do efeito passasse para 7,7% e 6,1% respectivamente em LP e MAT, o que reforça a importância da redução das desigualdades socioeconômicas para uma melhoria nos indicadores educacionais.

Os resultados obtidos estão alinhados com os estudos que associam maior uso da avaliação com melhoria do desempenho do objeto avaliado (Henry; Mark, 2003; Kirkhart, 2000; Mark; Henry, 2004) e representam o esforço para que mais pesquisas avancem na mensuração dessa relação, algo ainda carente na literatura (Daigneault, 2014). Em termos de construção da capacidade avaliativa, ainda que com tamanho do efeito relativamente baixo, tanto em LP quanto em MAT, os coeficientes positivos e significativos reforçam a ideia de que o uso da avaliação é fundamental para dar sustentação a uma política avaliativa. Neste sentido, o SAEB mostra-se ainda em desenvolvimento da sua ECB, não pela sua capacidade de conduzir a avaliação, mas por sua reduzida capacidade de usá-la e de transformar o desempenho das escolas, notadamente pelas redes municipais de educação, uma vez que não há ECB sem uso efetivo (Al Hudib; Cousins, 2022; Bourgeois; Cousins, 2013).

Porém, há de se considerar que os resultados encontrados são limitados e ainda necessitam de mais pesquisas, notadamente porque as variáveis de uso da avaliação não são medidas validadas reportadas na literatura. A variável EU-Escola, por exemplo, além de ser conceitualmente pobre, possui forte concentração no limite superior, o que indicaria que praticamente todos os diretores de escolas fazem uso da avaliação, seja para ajustar seus PPP ou para definir metas de alcance de indicadores. A variável EU-Rede, mesmo tendo maior heterogeneidade entre as redes, com algumas fazendo mais uso e outras menos, possui itens que são bastante distintos entre si, como fornecer subsídios para a formação de professores e criar critérios para remanejamento de diretores. Dessa forma, sugere-se que pesquisas futuras refaçam a análise de uso da avaliação e desempenho das escolas com a utilização de escalas validadas (Bourgeois *et al.*, 2016; Paixão; Rodriguez, 2023).

## 7. Conclusão

Os resultados indicam que quanto maior o uso da avaliação por parte de diretores de escolas e, principalmente, por parte de secretários municipais de educação, maior será o

desempenho em LP e MAT no 5º ano do Ensino Fundamental I. Além de prática, essa conclusão representa um esforço teórico importante, notadamente diante da carência de estudos quantitativos sobre uso da avaliação (Daigneault, 2014). Porém, com as medidas de uso da avaliação disponíveis no SAEB, o tamanho do efeito foi baixo, o que indica que outras variáveis podem explicar melhor o desempenho do que os usos da avaliação por parte de diretores e secretários. Uma delas é justamente o INSE, que ajudou a aumentar bastante o tamanho do efeito. O que já fora observado em pesquisas anteriores (Scheerens, 2016).

Em termos práticos, tem-se algumas recomendações. A primeira delas é que é mais importante concentrar os esforços para a melhoria das escolas através da ampliação dos usos da avaliação no nível da rede (meso). Entende-se que a rede tem mais capacidade para orientar o processo e direcionar os esforços para a melhor do desempenho das escolas a partir das considerações de uso do SAEB. Além disso, o desempenho das escolas está intrinsecamente associado ao nível socioeconômico da mesma. Portanto, o efeito no desempenho pode ser impulsionado se, além da ampliação nos usos da avaliação, houver esforços para a elevação do nível socioeconômico da escola. Por fim, o esforço de construção de um sistema avaliativo tão grande quanto o SAEB e que custa tanto para o país necessita dar mais ênfase aos usos dos resultados que ele mesmo gera. A união da capacidade de conduzir com a capacidade de usar avaliações constrói a capacidade avaliativa de um sistema (Al Hudib; Cousins, 2022), de forma a garantir as diretrizes determinadas pela política avaliativa (nível macro) da educação básica brasileira.

## Referências


ACREE, J.; CHOUINARD, J. A. Exploring Use and Influence in Culturally Responsive Approaches to Evaluation: A Review of the Empirical Literature. **American Journal of Evaluation**, [*s. l.*], v. 41, n. 2, p. 201–215, 2020.

AL HUDIB, H.; COUSINS, J. B. Understanding Evaluation Policy and Organizational Capacity for Evaluation: An Interview Study. **American Journal of Evaluation**, [*s. l.*], v. 43, n. 2, p. 234–254, 2022.

ALKIN, M. C.; KING, J. A. Definitions of Evaluation Use and Misuse, Evaluation Influence, and Factors Affecting Use. **American Journal of Evaluation**, [*s. l.*], v. 38, n. 3, p. 434–450, 2017.

ALKIN, M. C.; KING, J. A. The historical development of evaluation use. **American Journal of Evaluation**, [*s. l.*], v. 37, n. 4, p. 568–579, 2016.

ALKIN, M. C.; TAUT, S. M. Unbundling evaluation use. **Studies in Educational Evaluation**, [*s. l.*], v. 29, n. 1, p. 1–12, 2002.

BATES, D. *et al.* Fitting Linear Mixed-Effects Models Using lme4. **Journal of Statistical Software**, [*s. l.*], v. 67, n. 1, p. 1–48, 2015.

BECKER, T. E. Potential Problems in the Statistical Control of Variables in Organizational Research: A Qualitative Analysis With Recommendations. **Organizational Research Methods**, [*s. l.*], v. 8, n. 3, p. 274–289, 2005.

BERNERTH, J. B.; AGUINIS, H. A Critical Review and Best-Practice Recommendations for Control Variable Usage. **Personnel Psychology**, [*s. l.*], v. 69, n. 1, p. 229–283, 2016.

BOURGEOIS, I. *et al.* Measuring Evaluation Capacity in Ontario Public Health Units. **Canadian Journal of Program Evaluation**, [*s. l.*], v. 31, n. 2, p. 165–183, 2016.

BOURGEOIS, I.; COUSINS, J. B. Informing Evaluation Capacity Building Through Profiling Organizational Capacity for Evaluation: An Empirical Examination of four Canadian Federal Government Organizations. **Canadian Journal of Program Evaluation**, [*s. l.*], v. 23, n. 3, p. 127–146, 2009.

BOURGEOIS, I.; COUSINS, J. B. Understanding Dimensions of Organizational Evaluation Capacity. **American Journal of Evaluation**, [*s. l.*], v. 34, n. 3, p. 299–319, 2013.

BRASIL. **Detalhamento da população e resultados do SAEB 2021**. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2021.

BRASIL. **Escalas de proficiência do SAEB**. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2020.

BRASIL. **Saeb 2021: Indicador de Nível Socioeconômico do Saeb 2021 - nota técnica**. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2023.

BROOKE, N.; SOARES, J. F. **Pesquisa em eficácia escolar: origem e trajetórias**. Belo Horizonte: Editora UFMG, 2008.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel Inference: Understanding AIC and BIC in Model Selection. **Sociological Methods & Research**, [*s. l.*], v. 33, n. 2, p. 261–304, 2004.

CHEN, D.-G.; CHEN, J. K. **Statistical Regression Modeling with R: Longitudinal and Multi-level Modeling**. Cham: Springer International Publishing, 2021. (Emerging Topics in Statistics and Biostatistics). Disponível em: https://link.springer.com/10.1007/978-3-030-67583-7. Acesso em: 28 jun. 2024.

DAIGNEAULT, P.-M. Taking stock of four decades of quantitative research on stakeholder participation and evaluation use: A systematic map. **Evaluation and Program Planning**, [*s. l.*], v. 45, p. 171–181, 2014.

DILLMAN, L. M.; CHRISTIE, C. A. Evaluation Policy in a Nonprofit Foundation: A Case Study Exploration of the Robert Wood Johnson Foundation. **American Journal of Evaluation**, [*s. l.*], v. 38, n. 1, p. 60–79, 2017.

FRYER, R. G. **Learning from the successes and failures of charter schools**. Cambridge, MA: The Education Innovation Laboratory, Harvard University, 2012.

GUSMÃO, F. A. F.; AMORIM, S. S. Revisão sistemática: pesquisas empíricas acerca do Saeb (1995-2021). **Estudos em Avaliação Educacional**, [*s. l.*], v. 34, p. e09051, 2023.

HENRY, G. T.; MARK, M. M. Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions. **American Journal of Evaluation**, [*s. l.*], v. 24, n. 3, p. 293, 2003.

KHINE, M. S. (org.). **Methodology for Multilevel Modeling in Educational Research: Concepts and Applications**. Singapore: Springer Singapore, 2022. Disponível em: https://link.springer.com/10.1007/978-981-16-9142-3. Acesso em: 28 jun. 2024.

KIRKHART, K. E. Reconceptualizing evaluation use: An integrated theory of influence. **New Directions for Evaluation**, [*s. l.*], v. 2000, n. 88, p. 5–23, 2000.

KUZNETSOVA, A.; BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. **Journal of Statistical Software**, [*s. l.*], v. 82, n. 13, p. 1–26, 2017.

LEEUW, F. L.; FURUBO, J.-E. Evaluation systems: what are they and why study them?. **Evaluation**, [*s. l.*], v. 14, n. 2, p. 157–169, 2008.

LÜDECKE, D. *et al.* performance: An R Package for Assessment, Comparison and Testing of Statistical Models. **Journal of Open Source Software**, [*s. l.*], v. 6, n. 60, p. 3139, 2021.

LÜDECKE, D. **sjPlot: Data Visualization for Statistics in Social Science**. Versão R package version 2.8.16. [*S. l.: s. n.*], 2024. Disponível em: https://CRAN.R-project.org/package=sjPlot.

LUO, W. *et al.* Reporting Practice in Multilevel Modeling: A Revisit After 10 Years. **Review of Educational Research**, [*s. l.*], v. 91, n. 3, p. 311–355, 2021.

MARK, M. M.; COOKSY, L. J.; TROCHIM, W. M. K. Evaluation policy: An introduction and overview. **New Directions for Evaluation**, [*s. l.*], v. 2009, n. 123, p. 3–11, 2009.

MARK, M. M.; HENRY, G. T. The Mechanisms and Outcomes of Evaluation Influence. **Evaluation**, [*s. l.*], v. 10, n. 1, p. 35–57, 2004.

NAKAGAWA, S.; JOHNSON, P. C. D.; SCHIELZETH, H. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. **Journal of The Royal Society Interface**, [*s. l.*], v. 14, n. 134, p. 20170213, 2017.

NEUMAN, A. *et al.* Evaluation utilization research—Developing a theory and putting it to use. **Evaluation and Program Planning**, [*s. l.*], v. 36, n. 1, p. 64–70, 2013.

PAIXÃO, R. B.; RODRIGUEZ, M. C. Development and preliminary validation of the evaluation use scale for evaluation systems (EUS-ES). **Educational Research and Evaluation**, [*s. l.*], v. 28, n. 4–6, p. 100–115, 2023.

PRESKILL, H.; BOYLE, S. A Multidisciplinary Model of Evaluation Capacity Building. **American Journal of Evaluation**, [*s. l.*], v. 29, n. 4, p. 443–459, 2008.

RAUDENBUSH, S. W.; BRYK, A. S. **Hierarchical Linear Models: Applications and Data Analysis Methods**. 2. ed. Thousand Oaks: SAGE Publications, 2002.

SCHEERENS, J. **Educational Effectiveness and Ineffectiveness**. Dordrecht: Springer Netherlands, 2016. Disponível em: http://link.springer.com/10.1007/978-94-017-7459-8. Acesso em: 29 abr. 2025.

SPECTOR, P. E.; BRANNICK, M. T. Methodological Urban Legends: The Misuse of Statistical Control Variables. **Organizational Research Methods**, [*s. l.*], v. 14, n. 2, p. 287–305, 2011.

STOCKDILL, S. H.; BAIZERMAN, M.; COMPTON, D. W. Toward a definition of the ECB process: A conversation with the ECB literature. **New Directions for Evaluation**, [*s. l.*], v. 2002, n. 93, p. 7–26, 2002.

TROCHIM, W. M. K. Evaluation policy and evaluation practice. **New Directions for Evaluation**, [*s. l.*], v. 2009, n. 123, p. 13–32, 2009.

VAN THIEL, S.; LEEUW, F. L. The performance paradox in the public sector. **Public Performance & Management Review**, [*s. l.*], v. 25, n. 3, p. 267–281, 2002.

YAREMYCH, H. E.; PREACHER, K. J.; HEDEKER, D. Centering categorical predictors in multilevel models: Best practices and interpretation. **Psychological Methods**, [*s. l.*], v. 28, n. 3, p. 613–630, 2023.