Modelos de valor agregado para medir a eficácia das escolas Geres[1]

Tufi Machado Soares [a]
Alicia Bonamino [b]
Nigel Brooke [c]
Neimar da Silva Fernandes [d]

Resumo

A partir da pesquisa longitudinal Geres, o estudo propõe a comparação de diferentes modelos estatísticos com graus variados de complexidade para determinar a eficácia de escolas de Ensino Fundamental. O propósito da comparação é o de determinar se um grau maior de complexidade se justifica em termos de maior precisão e se há diferenças entre os modelos na sua consistência e capacidade de retratar de modo estável o desempenho da escola. Os dois modelos mais simples da contribuição da escola para a proficiência do aluno, denominados modelos de *Status*, incorporam ou uma medida do nível socioeconômico médio da escola, ou uma medida da condição socioeconômica de cada aluno como *proxy* para a proficiência prévia. Os outros dois, denominados de modelos de Valor Agregado (VA), incorporam medidas da proficiência prévia, o que os tornam modelos em condições de descrever o ganho de aprendizagem atribuível à escola no período em estudo. O estudo indica alta correlação entre os modelos de VA, mas baixa correlação deles com os modelos de *Status*, mostrando que é pequeno o ganho de precisão com a adição de uma medida da condição socioeconômica de cada aluno. Descobre-se que cerca de 80% das escolas apresentam estabilidade para as diferentes medidas temporais de VA, sugerindo que a eficácia seja, de fato, uma característica razoavelmente estável no tempo, e que o VA pode contribuir para a comparação das escolas e a definição de intervenções, pelo menos no primeiro segmento do Ensino Fundamental.

Palavras-chave: Avaliação educacional. Valor agregado. Modelo multinível.

[a] Universidade Federal de Juiz de Fora – UFJF. Juiz de Fora, Minas Gerais, Brasil.

[b] Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio. Rio de Janeiro, Rio de Janeiro, Brasil.

[c] University of London. London, Reino Unido.

[d] Universidade Federal de Juiz de Fora – (UFJF). Juiz de Fora, Minas Gerais, Brasil.

[1] Os autores agradecem ao apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), por meio do Projeto "Rede Multinível de Pesquisa em Eficácia Escolar" (Casadinho/Procad) e à Capes e ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), por meio do Projeto "As Inconsistências na Aprendizagem de Leitura e Matemática nos Anos Iniciais do Ensino Fundamental" (Observatório da Educação), assim como ao Centro de Avaliação de Políticas Públicas da Educação da Universidade Federal de Juiz de Fora (CAEd/UFJF), que foram essenciais para a elaboração deste artigo.

Recebido em: 07 mai. 2015
Aceito em: 30 jun. 2016

1 Introdução

Tem sido crescente a demanda pela melhoria da qualidade dos serviços públicos. No entanto, definir a qualidade de forma mensurável para poder monitorar o avanço em direção aos níveis desejados tem se demonstrado uma tarefa bastante complexa. Uma das demandas importantes feita por gestores, em particular, mas por toda a sociedade, em geral, é um sistema de medição capaz de inferir sobre a qualidade da escola, fornecendo uma medida confiável do quanto uma escola contribui para o aprendizado do aluno que ela atende. Os usos para tal medida são diversos, e vão desde ser uma referência para a escolha da escola pelos pais até servir de apoio nas políticas de responsibilização *high stakes*.

Enquanto se aguarda esta medida, o Índice de Desenvolvimento da Educação Básica (IDEB) tem se convertido em sinônimo da qualidade da educação, ao ponto de instigar a colocação de placas contendo esta informação na porta das escolas de diversos sistemas estaduais de ensino, incluindo os de Goiás, Tocantins, Acre e Minas Gerais. O IDEB é calculado pelo produto de dois outros índices: um que mede a taxa de aprovação dos alunos e outro baseado na média das proficiências em Português e Matemática na Prova Brasil. A literatura aponta uma série de problemas quanto a esse tipo de informação, que é usado para comparar escolas, como se fornecesse uma medida de quanto uma escola é responsável pelo aprendizado dos seus alunos (SOARES; XAVIER, 2013). Os principais problemas apontados são o IDEB não considerar o perfil sociodemográfico do aluno, nem o contexto social da escola, além de não levar em consideração, no resultado final do aluno, o seu desempenho prévio no começo de cada etapa escolar. Ou seja, o indicador oferece algumas informações importantes, mas não é uma medida direta da eficácia da escola. Para avaliar diretamente a eficácia da escola, é preciso saber acerca dos antecedentes do aluno, e o quanto ele teria aprendido se tivesse estudado em uma escola média do mesmo sistema.

Pode-se contornar uma parte dos problemas apontados utilizando o que se acostumou denominar modelos de *Status* ou modelos de resultados contextualizados (FERRÃO, 2014), que consideram o contexto social da escola e o perfil sociodemográfico do aluno. Mesmo nesse caso, ainda persiste a última questão da proficiência prévia do aluno, necessária para medir o ganho de aprendizado atribuível ao trabalho da escola ao longo das etapas escolares. Ao acrescentar essa informação, criam-se as condições para a formulação de Modelos de Valor Agregado (VA) e do cálculo mais pormenorizado da eficácia da escola. Este avanço leva o gestor educacional mais próximo da mensuração de diferenças dos processos escolares.

Os Modelos de Valor Agregado referem-se, então, a uma família de modelos estatísticos que, quando adequadamente especificados, podem ser utilizados para inferir sobre a eficácia de escolas e professores. Assim, esses modelos têm adquirido destaque na pesquisa dos elementos associados a diferenças entre escolas e sistemas educacionais nos níveis de aprendizagem dos alunos. Além disso, cresce sua utilização por parte dos gestores educacionais, nos Estados Unidos e alguns países europeus, para a formulação de políticas de melhoria dos processos escolares, em políticas de responsabilização e, também, na avaliação de programas (FERRÃO; COUTO, 2013). Adicionalmente, no contexto do ensino privado e de sistemas públicos, que incentivam a escolha de escolas por parte dos pais, esses modelos têm sido empregados para melhor embasar a decisão da família sobre a escola em que pretende matricular os filhos. Entretanto, o uso precipitado de modelos VA, sem o reconhecimento necessário das dificuldades de medição e de estabilidade, tem provocado polêmica e resistência por parte de pais e professores (AMREIN-BEARDSLEY, 2014). Torna-se cada vez mais importante pesquisar os limites dos modelos VA e definir em quais condições eles podem ser empregados com o mínimo de risco. O objetivo dos modelos VA é estimar o quanto uma escola contribui para o crescimento daqueles conhecimentos e habilidades cognitivas dos seus alunos, que são passíveis de medição por meio de testes durante um determinado período escolar. A situação dos alunos pode ser levada em consideração por meio de comparações com a média de contribuições de todas as escolas. A média de contribuições de todas as escolas com alunos de perfis similares, dentro de um determinado sistema, também pode ser considerada. Assim, a concepção de valor agregado está fundamentalmente ligada à ideia de causalidade (RUBIN; STUART; ZANUTTO, 2004), no sentido de que são as diferenças nos processos escolares, portanto intrínsecos à escola, que causam as diferenças nas medidas de valor agregado.

Como, na prática, análises causais são difíceis de serem fundamentadas no ambiente educacional, e considerando que as alternativas conhecidas envolvem estudos experimentais, tipicamente com grupos de tratamento e controle, nos quais os indivíduos que são designados para um grupo ou outro são escolhidos aleatoriamente, são sugeridos estudos observacionais, ou *quasi*-experimentais, como alternativas viáveis. Neste caso, as análises são baseadas na construção de modelos estatísticos dos quais os modelos de VA são exemplos típicos. Reardon e Raudenbush (2009) apresentam várias hipóteses fundamentais, que devem ser satisfeitas para justificar a inferência causal a partir desses modelos. Não sem controvérsias, a literatura sugere que a inferência de causalidade pode ser estimada a partir dos resultados desses modelos, desde que sejam suficientemente complexos e/ou determinadas hipóteses sejam verificadas (BRAUN; WAINER, 2007).

Essa complexidade é necessária, a princípio, para assegurar que as unidades expostas a um tratamento particular sejam equivalentes – ou aproximadamente equivalentes – àquelas com as quais estão sendo comparadas. Há uma grande discussão na literatura sobre quão complexos devem ser esses modelos (BRAUN; CHUDOWSKY; KOENIG, 2010). Entre outros, é sugerido que se considerem os seguintes aspectos: observação longitudinal dos alunos; viés de escolha da escola (também conhecido como autosseleção), que pode, em parte, ser controlado pelo *status* socioeconômico da família; erro de medida dos testes; *missing data* – atrito entre a amostra de alunos presentes ao teste e o total de alunos da escola, ocasionado por diferentes fatores; desempenho e evolução do desempenho associado ao *status* socioeconômico familiar e à variabilidade das medidas de VA ao longo do tempo[2].

Apesar de apontar possíveis problemas com as medidas de VA, a literatura ainda avaliou muito pouco os efeitos de cada um desses fatores na consistência e validade dessas medidas, e muitos estudos concluíram que, para determinados contextos, esses efeitos são pouco apreciáveis (FELDMAN; RABE-HESKETH, 2012; REARDON; RAUDENBUSH, 2009). Por outro lado, em outros contextos, eles podem ser substanciais (FERRÃO; COUTO, 2013; GRAY et al., 1995; LECKIE; GOLDSTEIN, 2009;). Assim, entender o funcionamento das medidas de VA, obtidas por meio dos diferentes modelos, é uma contribuição importante para que o gestor possa escolher a alternativa mais adequada dentro da realidade dos sistemas de avaliação educacional. As medidas mais elaboradas precisam ser construídas por meio de estudos e modelos mais complexos que podem levar a custos mais altos sem, necessariamente, oferecerem resultados significativamente diferentes das medidas menos elaboradas.

O estudo de Ferrão (2014) aborda os problemas de escolha do modelo de VA e o das propriedades de consistência e estabilidade dos indicadores de VA produzidos no contexto educacional brasileiro, bem como as vantagens e limitações dos seus usos. Por consistência, entende-se a capacidade de diferentes modelos de produzirem resultados similares. Por estabilidade, procura-se a capacidade dos modelos de classificarem as escolas da mesma forma ao longo de um período de tempo, supondo-se que a eficácia da escola deva apresentar alguma estabilidade para uma atuação exitosa. Utilizando a aplicação de modelo multinível aos dados do Geres 2005, o estudo mostra que os indicadores produzidos pelos modelos de *Status* são muito diferentes dos de valor agregado, reafirmando a necessidade de adoção de estudos longitudinais e sugerindo a sua utilidade para diagnósticos de eficácia escolar. O trabalho de Ferrão (2014) conclui ainda que existe uma

[2] Para uma discussão mais detalhada ver, por exemplo, Braun, Chudowsky e Koenig (2010).

estabilidade considerável para as medidas de VA, o suficiente para dividir as escolas em dois grupos: um que ela denomina de grupo das escolas eficazes e o outro, grupo das escolas não eficazes.

Na continuidade do trabalho de Ferrão (2014), com dados Geres, consideramos três modelos de VA e três modelos de *Status*, procurando evidências para determinar se o modelo de *Status* pode também ser usado para identificar a eficácia da escola, e se algum dos modelos construídos de VA pode ser considerado mais apropriado para essa finalidade do que outro. No entanto, nossos modelos de *Status* se distinguem do modelo empregado por Ferrão (2014), ao considerarem não só a condição socioeconômica dos alunos, mas também a condição socioeconômica média dos alunos da escola. Além disso, propomos testar a estabilidade das medidas de VA em quatro níveis diferentes de eficácia que, se bem-sucedidas, permitiriam maior flexibilidade ao gestor, por exemplo, na calibração de uma política de intervenção apropriada para as escolas de cada nível.

Além desta Introdução, o artigo está organizado em quatro seções. Na seção seguinte, é apresentada a pesquisa Geres, cujos dados são usados para estimar os modelos de VA e o de *Status*. A seção três apresenta os diferentes modelos de VA empregados na avaliação da eficácia das escolas participantes do estudo longitudinal. Na seção quatro, apresentam-se os resultados comparativos entre os diferentes modelos de eficácia e a análise de estabilidade das medidas de VA para um dos modelos estudados. Por último, são feitas algumas considerações sobre os resultados.

2 Geres – Estudo longitudinal da Geração Escolar 2005

O Geres foi uma pesquisa longitudinal, que focalizou a aprendizagem nos primeiros anos do Ensino Fundamental para estudar os fatores escolares e sociofamiliares que incidem sobre o desempenho e equidade escolar (BROOKE, BONAMINO, 2011). Durante um período de quatro anos, de 2005 a 2008, mais de 21 mil alunos, de uma amostra de 303 escolas estaduais, municipais e particulares, foram testados todo ano, em Língua Portuguesa e Matemática, enquanto os professores, diretores, pais e os próprios alunos foram entrevistados para determinar os impactos na aprendizagem dos fatores escolares e familiares. A primeira onda de medida foi tomada em março de 2005, com a observação de alunos da 1ª série (2º ano) do Ensino Fundamental (ou seu equivalente, quando a organização do ensino era em ciclos), e a segunda onda de medida ocorreu em outubro/novembro do mesmo ano. O painel foi observado também em novembro de 2006, 2007 e 2008, viabilizando o acompanhamento da amostra de alunos ao longo de quatro anos letivos. As escolas da amostra estão localizadas em Belo Horizonte, Rio

de Janeiro, Campinas, Campo Grande e Salvador, e as seis universidades que participaram do planejamento, coordenação e execução do projeto foram UFMG, PUC-Rio, UEMS, UFBA, Unicamp e UFJF.

Uma vez selecionadas as escolas, foram incorporados pelo Geres todos os alunos da 1ª série que já tinham alguma exposição prévia à alfabetização, visto que a metodologia da pesquisa, em termos do tipo de prova a ser aplicada na 1ª onda, pedia o nível de compreensão do código de escrita típico de alunos com um ano de alfabetização. Na prática, este critério permitiu a incorporação de todos os alunos que tinham passado pela pré-escola ou por classes de alfabetização. Como a 1ª série representava a primeira exposição do aluno ao processo de alfabetização na grande maioria das escolas da cidade de Salvador, e algumas de Campinas, nessas escolas, foi necessário iniciar a pesquisa com alunos da 2ª série, que hoje corresponde ao 3º ano do Ensino Fundamental de nove anos. Para elas, a pesquisa Geres se encerrou em 2007, um ano antes do encerramento da coleta de dados nas demais.

Dos 21.529 alunos cadastrados como "Alunos Geres", no primeiro semestre de 2005, quase metade já não se encontrava mais em escolas da pesquisa no segundo semestre de 2008. Uma parte dessa perda se deve ao processo constante de transferência de alunos de uma escola para outra, inclusive de outras redes e cidades. A maior parte da perda, no entanto, decorreu da retirada da cidade de Salvador, um ano antes do final da pesquisa, o que acarretou a perda do último ano do painel naquela cidade. Por outro lado, os alunos reprovados foram também avaliados e acompanhados na pesquisa. Com essas perdas, o grupo de alunos que estiveram presentes nas escolas da pesquisa, até o final de 2008, é composto de 10.836 casos. Ou seja, dos 21.529 alunos cadastrados nas escolas selecionadas no primeiro ano, apenas 10.836 (50,3%) chegaram a se cadastrar de novo na onda final da pesquisa.

A definição efetivamente empregada para a Geração Escolar 2005 foi, então, a presença do aluno no primeiro e último recadastramento dos alunos, efetuado todo ano, no princípio do 2º semestre. O aluno que esteve presente na escola ao longo desse período, independentemente do número de vezes em que fez as provas, pode ser considerado membro da Geração Escolar 2005.

O Geres buscou também agregar informações a respeito dos pais e/ou responsáveis dos alunos para determinar o nível socioeconômico de todos os alunos participantes, o que tem se tornado variável fundamental para a compreensão do percurso de aprendizagem dos alunos. O questionário abrangia questões sobre escolaridade,

presença de itens de conforto na residência, endereço, principal ocupação e funções desempenhadas no trabalho pelos pais ou responsáveis dos alunos. A metodologia empregada para o cálculo de nível socioeconômico (NSE) levou em consideração as questões referentes aos seguintes construtos: Escolaridade, Renda e Ocupação. Os construtos, as variáveis e as categorias presentes no questionário de pais e utilizadas no cálculo estão descritas na Tabela 1.

Nota-se que o método levou em conta apenas a escolaridade da mãe, descartando a informação sobre o pai. Essa escolha deveu-se ao fato de o efeito da mãe, na escolarização dos filhos, ser considerado superior ao do pai (BUCHMANN, 2002). Ao final do processo de cálculo, foi construída uma base com quatro variáveis: as duas informações de Escolaridade da mãe, o índice estimado de Renda e o índice que representa a Ocupação de maior *status* socioeconômico entre os pais. Estas variáveis foram levadas ao Multilog, que estimou o NSE de todos os alunos participantes.

A Tabela 2 revela que o NSE médio da Geração Escolar 2005 se situa ligeiramente acima da média do grupo de alunos participantes como um todo, por não incluir alunos dos níveis socioeconômicos mais baixos. Por sua vez, o Gráfico mostra que os alunos de nível socioeconômico mais baixo se concentram nas escolas públicas, e que os alunos das escolas do estrato especial[3] e estrato privado são basicamente os mesmos em termos socioeconômicos.

Para as análises produzidas neste trabalho foram consideradas apenas as escolas que possuíam pelo menos 15 alunos da Geração Escolar 2005, o que reduziu o total da amostra para 172 escolas e 7.391 alunos. Com esta escolha, pretendeu-se minimizar os efeitos dos erros de medidas, assim como possíveis efeitos amostrais sobre as médias de proficiências das escolas.

As medidas de proficiências foram calculadas pelo uso da Teoria da Resposta ao Item, para as disciplinas de Matemática e Língua Portuguesa, equalizadas ao longo dos anos pelo método de calibração simultânea considerando múltiplos grupos.

3 Modelos de *Status* e modelos de VA

3.1 A importância de comparar diferentes modelos de VA

O objetivo deste trabalho é apresentar uma análise comparativa de diferentes modelos de VA empregados na avaliação da eficácia das escolas participantes do

[3] Composto de escolas federais e outras escolas públicas ligadas a instituições de Ensino Superior, em Rio de Janeiro e Belo Horizonte.

Tabela 1. Construtos, itens e categorias da escala de nível socioeconômico.

Construto	Variável	Categorias
Escolaridade	Escolaridade da mãe ou responsável do sexo feminino	1 = Nunca estudou ou não chegou a terminar a 4ª série 2 = Terminou a 4ª série 3 = Terminou a 8ª série 4 = Terminou o Ensino Médio 5 = Terminou a faculdade
Renda (itens de conforto)	TV por assinatura	1 = Não tem 2 = Tem
	Máquina de lavar	1 = Não tem 2 = Tem
	Empregada doméstica	1 = Não tem 2 = Tem uma ou mais
	Sala	1 = Não tem 2 = Tem uma ou mais 3 = Tem duas ou mais
	Banheiro	1 = Não tem ou tem um 2 = Tem dois ou mais
	TV em cores	1 = Não tem 2 = Tem uma 3 = Tem duas ou mais
	Vídeo	1 = Não tem 2 = Tem uma ou mais
	Geladeira	1 = Não tem 2 = Tem uma ou mais
	DVD	1 = Não tem 2 = Tem uma ou mais
	Computador	1 = Não tem 2 = Tem uma ou mais
	Telefone fixo	1 = Não tem 2 = Tem uma ou mais
	Telefone celular	1 = Não tem 2 = Tem um 3 = Tem dois ou mais
	Carro	1 = Não tem 2 = Tem um 3 = Tem dois ou mais
Ocupação	Principal ocupação da mãe ou responsável do sexo feminino	
	Atividades realizadas na principal ocupação da mãe ou responsável do sexo feminino	
	Principal ocupação do pai ou responsável do sexo masculino	
	Atividades realizadas na principal ocupação do pai ou responsável do sexo masculino	

Fonte: BROOKE; BONAMINO, 2011.

estudo longitudinal Geres, analisando a correlação entre essas diferentes medidas e sua estabilidade. O primeiro tipo de estudo é chamado de análise de consistência de modelos de VA (ver FERRÃO; COUTO, 2013; GRAY et al., 1995). Esse estudo objetiva a comparação das medidas de eficácia escolar produzidas por diferentes modelos, avaliando a necessidade ou não de se empregar modelos mais complexos, com maior estrutura funcional, e analisando a necessidade de considerar outras variáveis de alunos e escolas na construção de medidas de eficácia escolar.

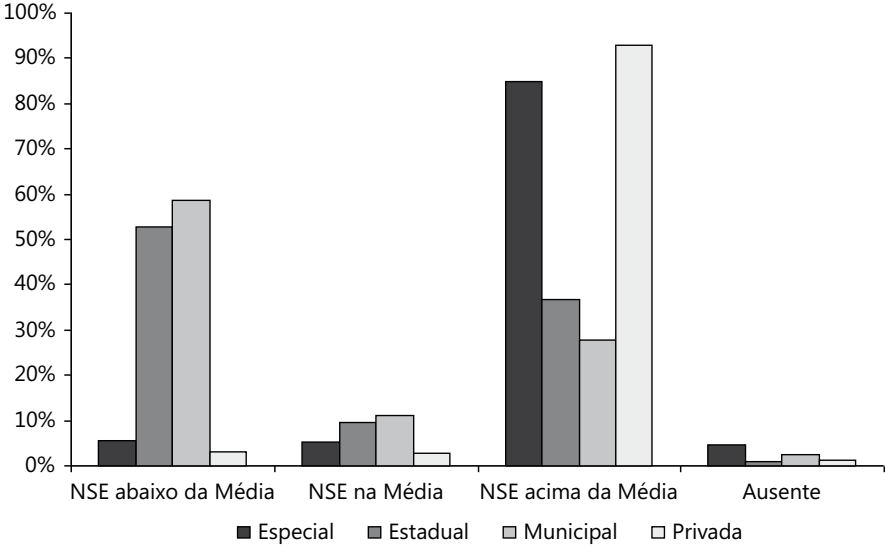
Por exemplo, Tekwe et al. (2004) produziram uma comparação de quatro modelos de VA utilizando três coortes de alunos, 2º para o 3º ano, 3º para o 4º ano e 4º para o 5º ano, para um distrito escolar de tamanho médio, na Flórida, com 22 escolas.

Tabela 2. Perfil das faixas de NSE segundo população.

		Total de participantes	Total com NSE	Mínimo	Mínimo	Média	Desvio padrão
NSE	Alunos cadastrados	39.342	28.445	-1,661	1,661	0,03223	0,556757
	Geração Escolar 2005	10.836	10.622	-1,453	1,661	0,07010	0,605292

Fonte: BROOKE; BONAMINO, 2011.

Gráfico. Nível socioeconômico Geração Escolar 2005, por estrato amostral de escolas.



Fonte: BROOKE; BONAMINO, 2011.

Foram obtidas medidas consecutivas em Matemática e leitura no *Iowa Test of Basic Skills* (ITBS), em 1989 e 1999. Os modelos variaram entre simples e complexos, e suas principais conclusões sugerem que os modelos que consideram estruturas multivariadas, para as disciplinas avaliadas, apresentaram correlações altas com modelos mais simples, por exemplo, os que consideram estruturas univariadas para o cálculo do valor agregado. Os autores concluíram que não parece haver vantagem substancial no uso de modelos mais complexos.

Por outro lado, observaram que determinadas variáveis sociodemográficas e escolares afetavam substancialmente as medidas de VA, no sentido de que existiam correlações menos expressivas entre as medidas de VA produzidas pelos modelos que consideravam essas variáveis e os que não as consideravam. Os autores, no entanto, observaram que a decisão de considerar, por exemplo, um indicador da condição socioeconômica em um modelo onde o nível de proficiência já é considerado na entrada é mais uma decisão política do que técnica, tendo em vista que levar em consideração a medida da condição socioeconômica poderia isentar a escola de sua responsabilidade para com o perfil de aluno que a procura.

No entanto, outros autores, como Braun, Chudowsky e Koenig (2010), afirmam que o crescimento nos níveis de proficiências depende também da condição econômica familiar do aluno, para além do esforço escolar. Essa escolha remete a um modelo mais complexo no cômputo da medida de VA da escola do que o utilizado habitualmente, como será visto na seção seguinte.

A flutuação ou instabilidade das medidas de VA é apontada na literatura como um importante obstáculo na avaliação da eficácia escolar, principalmente para programas de responsabilização com consequências fortes (*high-stakes)*. De fato, não é possível haver decisão em um sistema no qual as medidas de qualidade variam aleatoriamente. É particularmente crítica na avaliação de escolas pequenas, e quando se deseja avaliar o trabalho docente, mas, mesmo para as escolas grandes, a flutuação das medidas de VA pode ser razoavelmente perceptível.

Por exemplo, Ballou (2005) estudou a estabilidade dos ranqueamentos de professores derivados do modelo de VA empregado no Estado norte-americano do Tennessee, entre 1998 e 1999, para professores de Ensino Básico (*Elementary*) e de Ensino Médio (*Middle*), em um distrito de tamanho médio no Estado. Ele encontrou que 40% dos professores de Matemática que se encontravam no primeiro quartil (mais baixo) no ranqueamento, em 1998, continuavam no primeiro quartil, em 1999; no entanto, 30% deles estavam acima da mediana, em 1999. A estabilidade foi maior para os professores que se encontravam no quarto

quartil (mais alto), em 1998; apesar disso, quase um quarto deles se encontrava abaixo da mediana, em 1999.

Braun, Chudowsky e Koenig (2010) apontam que tal flutuação pode ser devida a erros de medida, mas, também, a outras fontes naturais de variação, tais como trocas na forma de ensinar de um ano para o outro. Como um alto nível de instabilidade é um problema para o uso do efeito estimado do professor em sistemas de *accountability high-stakes*, os autores sugeriram usar uma média de estimativas de valor agregado de três anos (*three year rolling average*), como forma de estabilizar as medidas de VA. Ferrão e Couto (2013), seguindo Gray et al. (1995), denominam esse tipo de estudo análise de estabilidade das medidas de VA.

É evidente, portanto, que não se deseja um sistema totalmente estável, pois isso implicaria em dizer que as políticas educacionais não têm efeitos, e que as ações internas das escolas não produzem melhoras ou pioras no aprendizado. A questão da análise da estabilidade é verificar a existência de um contexto estável o suficiente para a decisão, por exemplo, de diferenciar grupos de escolas de acordo com suas virtudes e necessidades, de modo a definir o tipo de intervenção apropriada a cada uma delas.

De fato, deve-se ter em mente que as medidas de VA são apenas um elemento em um sistema de decisão, e que devem ser acompanhadas de outros estudos e medidas de contexto para serem de fato úteis no apoio à decisão dos gestores.

Neste trabalho, seis modelos foram considerados: três modelos de *Status* e três modelos de VA. Trataremos primeiro dos modelos de *Status*.

3.2 Modelos de *Status*

Nos modelos *Status,* ou deposição (*standing*), é considerada apenas uma medida da proficiência ao final de uma etapa escolar, ainda que se considerem outros controles, por exemplo, a condição econômica média dos alunos – o nome *Status* deriva dos modelos de classificação nos quais os alunos são classificados em níveis ou *status* de aprendizado. O primeiro deles é o modelo de *Status* mais simples, conhecido como modelo nulo na literatura dos modelos multiníveis – cf. Raudenbush e Bryk (2002):

Modelo 0 (nulo) – Modelo de *Status* simples sem contextualização;

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$$
$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

No equacionamento acima, Y_{ij} representa uma variável de desfecho ou resultado – medida em uma escala contínua, tipicamente a proficiência do aluno *i* da escola *j*. O termo β_{0j} representa a contribuição da escola *j* para a proficiência do aluno, e γ_{00}, a contribuição média de todas as escolas. As hipóteses básicas sobre as componentes de incerteza são: $\varepsilon_{ij} \sim N(0,\sigma_{\varepsilon}^2)$, $\mu_{0j} \sim N(0,\sigma_u^2)$, e independentes para todo *i* e *j*. A medida da eficácia escolar é, então, dada por μ_{0j}.

O segundo deles é um modelo de *Status* também simples, mas com controle pela condição socioeconômica média da escola[4]:

Modelo 1 – Modelo de *Status* simples contextualizado;

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}\, NSE_j + \mu_{0j}$$

Neste caso, ele deve ser empregado se for desejo do gestor ou analista considerar o efeito médio da condição socioeconômica na medida de eficácia da escola *j*. Esta é representada, também, pelo termo μ_{0j}.

Esse modelo, apesar de ser muito simples, talvez seja o único possível de ser empregado, direta ou indiretamente, por exemplo, em algumas das políticas atuais de responsabilização no Brasil. Isso ocorre porque boa parte das informações disponíveis sobre os níveis de proficiências cognitivas dos alunos para as escolas brasileiras são obtidas por meio de avaliações seccionais, como a Prova Brasil, nas quais a condição socioeconômica dos alunos é obtida por meio de questionários contextuais aplicados simultaneamente e que, portanto, podem ser agrupadas para produzir uma medida da condição econômica da escola. Nesse modelo, obviamente, não há como avaliar o ganho agregado, e admite-se, implicitamente, que a proficiência avaliada seccionalmente traga consigo todo o efeito de eficácia da escola. É possível incluir outras variáveis do contexto escolar no modelo, mas não há consenso entre especialistas sobre quais variáveis (se alguma) deveriam ser usadas. Isso será mais discutido nas seções subsequentes, mas, num primeiro momento, poderíamos especular que a inclusão de uma determinada variável de processo escolar descontaria justamente o efeito que a medida de valor agregado deveria incluir. Exemplificando, é importante sabermos que a existência de biblioteca afeta a eficácia da escola, mas seria injusto descontarmos esse efeito, a menos que estivéssemos buscando aquelas escolas mais eficazes, mesmo nas situações mais desfavoráveis, isto é, sem bibliotecas. Da mesma forma,

[4] Ferrão e Couto 2013 utilizam o termo Modelo de Resultados Contextualizados.

poderíamos especular que o NSE médio da escola é *proxy* de muitos processos escolares e que, portanto, incluí-lo no modelo seria retirar da medida de eficácia parte daquilo que gostaríamos de medir.

O terceiro modelo considerado é também um modelo de *Status.* No entanto, esse modelo tenta emular um modelo de VA introduzindo uma medida da condição socioeconômica do aluno como *proxy* da proficiência prévia.

Modelo 2 – Modelo de *Status* com NSE do aluno como *proxy* da proficiência prévia.

$$Y_{ij} = \beta_{0j} + \beta_1 NSE_{ij} + \varepsilon_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} NSE_j + \mu_{0j}$$

Nesse caso, a medida de eficácia da escola que emula uma medida de valor agregado é também representada pelo termo μ_{0j}. Outras covariáveis podem ser introduzidas no nível do aluno com a mesma finalidade, isto é, como *proxies* da proficiência prévia, por exemplo, o gênero, a escolaridade dos pais etc.

3.3 Modelos de VA

Em geral, a literatura recomenda que, para isolar o efeito da escola ou do programa, ao menos duas medidas ao longo do tempo devem ser consideradas. Recomenda-se também que outras medidas de contexto escolar, tais como medidas de *status* econômico e *background* familiar (BRAUN; CHUDOWSKY; KOENIG, 2010), e variáveis de processos escolares, tais como qualidade da liderança escolar, deveriam ser usados no cômputo da medida de VA. De qualquer forma, deve-se avaliar alguma medida de crescimento. Trata-se de uma contraposição aos modelos de *Status*.

Assim, entre os modelos de VA, o modelo 3 é o modelo de VA mais simples:

Modelo 3 – Modelo de VA com efeito aleatório somente no intercepto sem covariáveis explicativas.

$$Y_{ij}(t) = \beta_{0j} + \beta_1 Y_{ij}(t - 1) + \varepsilon_{ij}$$
$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

O termo $Y_{ij}(t - 1)$ representa uma medida de proficiência do aluno em um momento anterior; tipicamente, t representa uma observação ao final de uma etapa escolar

e t -1 o início dessa etapa ou final da etapa anterior. Note-se que o termo β_{0j} representa a contribuição da escola na proficiência do aluno e, portanto, a medida de VA da escola é dada por μ_{0j}.

Modelo 4 – Modelo de VA com efeito aleatório somente no intercepto com covariáveis.

$$Y_{ij}(t) = \beta_{0j} + \beta_1 Y_{ij}\,(t - 1) + \beta_2\, NSE_{ij} + \varepsilon_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}\, NSE_j + \mu_{0j}$$

Note-se que o termo β_{0j} representa a contribuição da escola na proficiência do aluno, agora, condicionado à proficiência em t -1 e à condição socioeconômica do aluno (representada pelo termo NSE_{ij}) sendo, portanto, também uma medida da contribuição da escola sobre o ganho de proficiência. Admite-se, ainda, que as componentes de incerteza (ou erro) sejam tais que: $\varepsilon_{ij} \sim N(0,\sigma^2_\varepsilon)$, $\mu_{0j} \sim N(0,\sigma^2_u)$, são independentes para todo i e j. Como nos modelos anteriores, a medida de VA da escola é dada por μ_{0j}.

Finalmente, o modelo 5, o mais complexo, permite a análise do VA da escola para diferentes perfis de alunos, conforme sua medida de proficiência na entrada da etapa escolar.

Modelo 5 – Modelo de VA com efeito aleatório no intercepto e no coeficiente linear.

$$Y_{ij}(t) = \beta_{0j} + \beta_{1j}\, Y_{ij}\,(t - 1) + \beta_2\, NSE_{ij} + \varepsilon_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{10}\, NSE_j + \mu_{0j}$$
$$\beta_{1j} = \gamma_{01} + \gamma_{11}\, NSE_j + \mu_{1j}$$

Nesse caso, o efeito escola sobre o aluno i é dado por:

$$VA_{ij} = \mu_{0j} + \mu_{1j} * Y_{ij}\,(t - 1)$$

Note-se que, segundo este modelo, as escolas que apresentam $\mu_{1j} = 0$ produzem um ganho uniforme médio para todos os alunos, independente da proficiência anterior. Por outro lado, para o caso de proficiências positivas – portanto não centralizadas em torno da média do grupo – o que ocorre neste trabalho, se $\mu_{1j} > 0$ então o ganho é maior para os alunos com maiores níveis de proficiência na entrada, aumentando, portanto, a desigualdade ao final da etapa escolar considerada. E, se $\mu_{1j} < 0$ o ganho é

maior para os alunos com menores níveis de proficiência, diminuindo a desigualdade em relação às proficiências no tempo anterior. Admite-se que haja correlação entre os termos μ_{0j} e μ_{1j}. A medida de VA da escola é calculada, então, considerando o efeito médio sobre todos os alunos, isto é: $VA_j = \mu_{0j} + \mu_{1j} * E\,[Y_{ij}(t - 1)]$, onde $E\,[Y_{ij}\,(t - 1)]$ representa a média de proficiências em t - 1.

4 Análises de estabilidade e consistência das medidas de eficácia

Nesta seção, apresentam-se os resultados dos dois estudos realizados. O primeiro trata de uma comparação entre as medidas de eficácia escolar obtidas tanto por meio dos modelos de VA quanto por meio dos modelos de *Status*. O objetivo é verificar se os modelos mais complexos apresentam resultados significativamente diferentes dos modelos mais simples. Particularmente, pretende-se comparar os resultados dos modelos de VA com os obtidos pelos modelos de *Status*. Na seção 4.2, apresenta-se uma análise da estabilidade do modelo de VA 4. Os resultados para os demais modelos de VA são similares.

Para as análises envolvendo os modelos de VA, foram considerados os ganhos agregados (VA) entre duas diferentes ondas, sucessivas ou não: o valor agregado da escola na proficiência da onda 2 sobre a proficiência da onda 1 (2 predita por 1 ou, como grafado aqui, 2 pred 1), na proficiência da onda 3 sobre a proficiência da onda 2 (3 pred 2), onda 4 e onda 3 (4 pred 3), onda 5 e onda 4 (5 pred 4), onda 3 e onda 1 (3 pred 1), onda 5 e onda 3 (5 pred 3) e onda 5 e onda 1 (5 pred 1).

4.1 Comparação entre os modelos de VA e os modelos de *Status*

Diversos autores, principalmente os que advogam o uso de avaliações longitudinais, apontam que existe uma variedade de problemas com os modelos de posição ou *status*, mas o principal é que eles ignoram o impacto nas médias de proficiência dos vieses de autosseleção das escolas pelos pais (FERRÃO; COUTO, 2014). Além disso, ao apresentar apenas os níveis alcançados pelos alunos ao final de uma determinada etapa escolar, ignoram as diferenças no desempenho inicial dos alunos, na entrada do sistema. Por outro lado, os modelos de VA levam em consideração as trajetórias obtidas pelos alunos nos testes para todos os níveis de proficiência. Mas eles também apresentam limitações, especificadas em suas hipóteses, que precisam ser verificadas. Assim, não se pode excluir *a priori* o uso de modelos de *Status* apenas pelas críticas gerais, sem uma comparação com os modelos de VA.

Neste estudo, realizamos uma comparação entre as medidas de VA produzidas pelos modelos de 3 a 5 (considerando as ondas 5 pred 1) com as medidas de

VA produzidas pelos modelos de 0 a 2, apenas com o resultado da 5ª onda. As correlações de Pearson e Spearman (que consideram apenas o posto ou a ordem da medida) são reportadas nas Tabelas 3 e 4.

Evidentemente, como esperado, as correlações entre as medidas de VA produzidas pelos modelos de 3 a 5 são altas, em ambas as disciplinas. Esse resultado é relativamente bem conhecido da literatura, que aponta alta consistência de resultados entre diferentes tipos de modelos de VA. A correlação entre os modelos 1 e 2 de *Status* é praticamente 1, o que indica que pouco ou nenhuma informação adicional é obtida com a medida individual da condição econômica do aluno, dada a medida de condição econômica do grupo. Isto indica que se pode utilizar apenas a medida composicional, o NSE médio da escola, não havendo a necessidade

Tabela 3. Correlação (Pearson/Spearman) entre as Medidas de Eficácia dos Modelos – Língua Portuguesa – Geração Geres.

	Modelo 0 (5ª onda)	Modelo 1 (5ª onda)	Modelo 2 (5ª onda)	Modelo 4 (5 pred 1)	Modelo 5 (5 pred 1)
Modelo 0 (5ª onda)	1/1	0,518*/0,527*	0,518*/0,528*	0,331*/0,278*	0,326*/0,272*
Modelo 1 (5ª onda)	0,518*/0,527*	1*/1*	1*/1*	0,626*/0,540*	0,615*/0,527*
Modelo 2 (5ª onda)	0,518*/0,528*	1*/1*	1*/1*	0,627*/0,541*	0,615*/0,527*
Modelo 3 (5 pred 1)	0,666*/0,657*	0,517*/0,424*	0,517*/0,425*	0,886*/0,833*	0,876*/0,825*
Modelo 4 (5 pred 1)	0,331*/0,278*	0,626*/0,540*	0,627*/0,541*	1/1*	0,987*/0,980*
Modelo 5 (5 pred 1)	0,326*/0,272*	0,615*/0,527*	0,615*/0,527*	0,987*/0,980*	1*/1*

* Correlação é significativa ao nível de p ≤ 0,05.
Fonte: Elaborada pelos autores (2016).

Tabela 4. Correlação (Pearson/Spearman) entre as Medidas de Eficácia dos Modelos – Matemática – Geração Geres.

	Modelo 0 (5ª onda)	Modelo 1 (5ª onda)	Modelo 2 (5ª onda)	Modelo 4 (5 pred 1)	Modelo 5 (5 pred 1)
Modelo 0 (5ª onda)	1/1	0,486*/0,532*	0,486*/0,531*	0,402*/0,425*	0,396*/0,395*
Modelo 1 (5ª onda)	0,486*/0,532	1*/1*	1/1*	0,825*/0,774*	0,804*/0,748*
Modelo 2 (5ª onda)	0,486*/0,531	1*/1*	1/1*	0,826*/0,774*	0,805*/0,748*
Modelo 3(5 pred 1)	0,855*/0,876	0,626*/0,564*	0,626*/0,564*	0,777*/0,723*	0,772*/0,711*
Modelo 4 (5 pred 1)	0,402*/0,425	0,825*/0,774*	0,826*/0,774*	1/1*	0,990*/0,984*
Modelo 5 (5 pred 1)	0,396*/0,395	0,804*/0,748*	0,805*/0,748*	0,990*/0,984*	1*/1*

* Correlação é significativa ao nível de p ≤ 0,05.
Fonte: Elaborada pelos autores (2016).

de uma medida socioeconômica individual. Há de se investigar se o NSE médio de toda a escola, extraído do Censo Escolar, por exemplo, produziria resultados parecidos com o NSE médio da amostra de alunos avaliadas – que nem sempre está disponível. Por outro lado, o NSE do aluno não parece ser uma boa *proxy* da proficiência prévia do aluno, além do que já é explicado pelo NSE médio do grupo. Há, portanto, uma diferença substancial entre os resultados aqui encontrados e os encontrados por Ferrão (2014), tendo em vista que a autora considera um modelo de *Status* com controle pela condição socioeconômica do aluno, somente, e não pela condição socioeconômica média dos alunos da escola.

Em Língua Portuguesa, existe uma correlação média (0,5 a 0,6) entre as medidas de eficácia escolar que foram obtidas pelos modelos de VA e as que foram obtidas pelos modelos de *Status* (modelos 1 e 2). No caso da disciplina de Matemática, elas são altamente correlacionadas, em torno de 0,8. Ora, alta correlação não implica em dizer que os resultados são precisamente os mesmos, mas sugere que, admitindo-se que uma medida produzida por um modelo de VA seja superior em relação a uma medida produzida por um modelo de *Status*, não é um absurdo o uso de medidas de eficácia escolar obtidas nas análises seccionais no caso da Matemática, desde que se produza o controle pelo nível socioeconômico médio dos alunos – o mesmo não pode ser afirmado para Língua Portuguesa.

As correlações mais baixas entre os modelos que não incluem a condição socioeconômica média dos alunos (modelo 0 – *Status*, modelo 3 – VA) e os demais indicam que o efeito sobre a medida desta variável é bastante apreciável. Incluir ou não esta variável é certamente um problema ainda sem solução adequada na literatura, tendo em vista a possível correlação entre a condição socioeconômica média da escola e a sua eficácia (efeito conhecido como *endogeneidade*), que tem sido objeto de investigação por vários pesquisadores. Por outro lado, a inclusão da condição socioeconômica individual dos alunos pouco altera os resultados dos modelos de VA.

Finalmente, uma análise do ajustamento dos modelos por meio de estatísticas de *deviance*, AIC (Critério de Informação de Akaike) e BIC (Critério Bayesiano de Schwarz), reportada na Tabela 5, sugere que o modelo mais ajustado entre todos é o modelo 4 – de VA, para ambas as disciplinas.

4.2 Análise da Estabilidade das Medidas de VA para a Geração Geres

O estudo apresentado nesta seção amplia e complementa o estudo apresentado por Ferrão e Couto (2013). Naquele trabalho, os autores comparam a estabilidade das medidas de VA ao longo das sucessivas ondas do Geres, considerando como

Tabela 5. Análise do Ajuste dos Modelos.

Modelo	Graus de liberdade	Língua Portuguesa			Matemática		
		Deviance	AIC	BIC	Deviance	AIC	BIC
0	2	66414,01	66418,01	66424,30	80398,57	80402,57	80408,86
1	2	66185,86	66189,86	66196,15	80146,77	80150,77	80157,06
2	2	66025,27	66029,27	66035,56	79966,55	79970,55	79976,84
3	2	64078,68	64082,68	64088,97	77933,72	77937,72	77944,01
4	4	64033,80	64041,80	64054,39	77861,92	77869,92	77882,51
5	4	64151,88	64159,88	64172,47	78058,98	78066,98	78079,57

AIC: Critério de Informação de Akaike; BIC: Critério Bayesiano de Schwarz.
Fonte: Elaborada pelos autores (2016).

proficiência prévia a proficiência medida na onda imediatamente anterior. Aqui, ampliamos as alternativas ao levar em consideração outras possíveis combinações de desfecho e medida prévia.

Foi realizada uma análise da estabilidade das medidas de VA produzidas pelos modelos de VA (modelos de 3 a 5) apresentados anteriormente. Foram considerados os ganhos agregados (VA) entre duas diferentes ondas, sucessivas ou não: o valor agregado da escola na proficiência da onda 2 sobre a proficiência da onda 1 (2 predita por 1 ou, como grafado aqui, 2 pred 1), na proficiência da onda 3 sobre a proficiência da onda 2 (3 pred 2), onda 4 e onda 3 (4 pred 3), onda 5 e onda 4 (5 pred 4), onda 3 e onda 1 (3 pred 1), onda 5 e onda 3 (5 pred 3) e onda 5 e onda 1 (5 pred 1). Calcularam-se, então, as correlações de Pearson e Spearman entre as medidas de VA obtidas para todas essas combinações de proficiência subsequente e proficiência prévia.

Os resultados indicam que existe uma baixa correlação nas medidas de VA entre pares sucessivos e curtos de ondas, por exemplo, entre (2 pred 1) e (3 pred 2). No entanto, os valores dessas correlações aumentam substancialmente na medida em que se restringem as análises às escolas com mais alunos participantes. Na amostra aqui considerada, de 175 escolas, o número mínimo de alunos para a geração Geres foi de 15. Em uma análise paralela, restrita às 75 escolas com pelo menos 40 alunos, as correlações em Matemática entre as medidas de VA (2 pred 1) e (3 pred 2) se elevam para 0,569 (situando-se em uma faixa de média correlação) (SOARES, 2015). O motivo mais importante, mas não único, para isso, provavelmente, diz respeito à precisão da média de proficiências estimada da escola, que é maior quando se tem mais alunos comparecendo aos testes.

Por outro lado, a correlação aumenta também à medida que os pares de ondas se distanciam e/ou a distância entre a onda da proficiência prévia para a onda da proficiência subsequente é maior, podendo-se aludir a um possível efeito acumulativo ou melhora na medição, tendo em vista que o número de itens nos testes aumenta segundo as ondas do Geres. Um dos revisores deste artigo sugeriu ainda que "uma especulação é que por serem anos finais e, talvez, nos ciclos, anos de decisão de retenção ou não na série (ano escolar), haja maior preocupação com a aprendizagem, e outra especulação é que esses também são anos alvos de avaliações externas", o que também são explicações bastante plausíveis. Esse fenômeno não foi identificado por Ferrão e Couto (2013), dado que eles observaram os ganhos agregados apenas nas ondas sucessivas. Assim, as medidas de VA para o par (5 pred 1) são razoavelmente correlacionadas com todos os demais pares e altamente correlacionadas com os pares (5 pred 3) e (3 pred 1). Esses resultados são mostrados separadamente para Língua Portuguesa e Matemática nas Tabelas 6 e 7. São apresentadas as correlações de Pearson/Spearman somente para o modelo 4, mas esses resultados são similares aos obtidos para os demais, assim como para a base total – e não apenas a geração Geres. Por outro lado, reforçando o que dissemos anteriormente, as correlações aumentam substancialmente quando a análise se restringe à amostra das escolas com pelo menos 40 alunos da geração Geres – 60 escolas.

Esses resultados sugerem, grosso modo, duas possibilidades de interpretação, provavelmente complementares. Na primeira, a justificativa seria a de que os erros de medida decaem com a distância entre a onda da proficiência subsequente e a onda da proficiência prévia. Nesse caso, o menor erro de medida de VA seria verificável para as medidas construídas para o modelo que considera o par de ondas mais distantes (5 pred 1), (5 pred 3) e (3 pred 1), e as correlações mais altas decorrentes das análises seriam devidas aos menores erros nas medidas observadas. Mas é possível, também, que as escolas apresentem pouca padronização no ritmo do processo de ensino, algumas concentrando maiores esforços em determinados anos escolares do que em outros anos, tanto devido às experiências vivenciadas com relação às necessidades dos alunos quanto por escolha pedagógica, e a eficácia seria mesmo melhor avaliada após ciclos mais longos. Note-se que o intervalo de observação 3 pred 1 corresponde ao final do ciclo de alfabetização que, na escola pública, geralmente se estende até o 3º ano do Ensino Fundamental.

É mais difícil justificar o fato de que as correlações entre as medidas de valores agregados entre os pares de ondas mais próximos são menores do que para os pares de ondas mais distantes. Uma possível explicação seria baseada na suposição de que as escolas fazem um diagnóstico do aprendizado no período anterior, em um processo de retroalimentação (avaliação e ação pedagógica), e alteram o ritmo de ensino.

De qualquer forma, os resultados sugerem que é melhor usar medidas de VA com maiores distâncias entre os pares de ondas para analisar o progresso na escola. Depreende-se, assim, que é melhor utilizar, respectivamente, as medidas de VA calculadas para o ganho de proficiência entre a onda 1 e a onda 5, onda 1 e onda 3 e entre onda 3 e onda 5.

Finalmente, com o objetivo de auxiliarmos os gestores em uma possível intervenção, dividimos as medidas de valor agregado das escolas de acordo com os quatro quartis. Ou seja, classificamos as escolas em quatro grupos (1, 2, 3 e 4), desde o de menor valor agregado (BOTTOM 25%) até o de maior valor agregado (TOP 25%). Analisamos, então, a estabilidade da classificação das escolas nesses quatro grupos entre as ondas 3 pred 1, 5 pred 3 e 5 pred 1, simultaneamente, por meio de tabelas de cruzamentos organizados em três camadas. Na Tabela 8,

Tabela 6. Modelo 4 (Língua Portuguesa).

	2 pred 1	3 pred 2	4 pred 3	5 pred 4	5 pred 1	5 pred 3	3 pred 1
2 pred 1	1/1	-	-	-	-	-	-
3 pred 2	0,040/-0,047	1/1	-	-	-	-	-
4 pred 3	0,269*/0,234*	0,123/0,101	1/1	-	-	-	-
5 pred 4	0,256*/0,233*	0,341*/0,333*	0,120/0,108	1/1	-	-	-
5 pred 1	0,412*/0,372*	0,501*/0,455*	0,618*/0,581*	0,660*/0,585*	1/1	-	-
5 pred 3	0,224*/0,184*	0,157*/0,097*	0,689*/0,690*	0,740*/0,670*	0,838*/0,773*	1/1	-
3 pred 1	0,533*/0,506*	0,771*/0,749*	0,309*/0,259*	0,409*/0,369*	0,775*/0,743*	0,338*/0,251*	1/1

* Correlação é significativa ao nível de p ≤ 0,05, Correlação de Pearson/Correlação de Spearman.
Fonte: Elaborada pelos autores (2016).

Tabela 7. Modelo 4 (Matemática).

	2 pred 1	3 pred 2	4 pred 3	5 pred 4	5 pred 1	5 pred 3	3 pred 1
2 pred 1	1/1	-	-	-	-	-	-
3 pred 2	0,087/0,098	1/1	-	-	-	-	-
4 pred 3	0,211*/0,191*	0,137/0,171*	1/1	-	-	-	-
5 pred 4	0,302*/0,253*	0,178*/0,215*	0,028/-0,060	1/1	-	-	-
5 pred 1	0,483*/0,410*	0,530*/0,543*	0,493*/0,473*	0,687*/0,636*	1/1	-	-
5 pred 3	0,216*/0,159*	0,030/0,030	0,564*/0,563	0,762*/0,704*	0,772*/0,711*	1/1	-
3 pred 1	0,590*/0,557*	0,826*/0,827*	0,213*/0,222*	0,312*/0,287*	0,746*/0,710*	0,173*/0,109	1/1

* Correlação é significativa ao nível de p ≤ 0,05, Correlação de Pearson/Correlação de Spearman.
Fonte: Elaborada pelos autores (2016).

são apresentados os resultados para Língua Portuguesa. Os resultados devem ser analisados da seguinte forma, começando pelo alto à esquerda na Tabela 8: 18 escolas foram classificadas no 1º quartil nos modelos construídos para os três pares de ondas, sete escolas foram classificadas no 1º quartil pelos modelos construídos para os pares de ondas 5 pred 1 e 5 pred 3 e 2º quartil para o par 3 pred 1; 9 escolas foram classificadas no 1º quartil para os pares 5 pred 1 e 3 pred 1 e 2º quartil pelo par 5 pred 3; e 1 escola foi classificada no 1º quartil pelo par 5 pred 1 e no 2º quartil pelos pares 5 pred 3 e 3 pred 1. Notamos coerência e razoável estabilidade na classificação desse conjunto de escolas representado na tabela por G1 pelos modelos no sentido de pouco deslocamento da escola de um quartil para o quartil vizinho, em função das diferentes medidas de VA. Claro que é apenas uma observação descritiva, mas útil aqui para efeito de análise. O critério

Tabela 8. Estabilidade da classificação das escolas para o Modelo 3 (VA) segundo os quatro quartis para as ondas 5 pred 3, 3 pred 1 e 5 pred 1 – Língua Portuguesa.

Quartil VA Onda 5 pred 1			Onda 3 pred 1				Total
			1	2	3	4	
1	Onda 5 pred 3	1	18 **G1**	7	(3)	0	28
		2	9	1	0	0	10
		3	(5)	0	0	0	5
		4	0	0	0	0	0
	Total						43
2	Onda 5 pred 3	1	0	1 **G2**	6	(3)	10
		2	0	9	6	(2)	17
		3	4	7	0	0	11
		4	(4)	(1)	0	0	5
	Total						43
3	Onda 5 pred 3	1	0	0	0	(4)	4
		2	0	1	5	8	14
		3	(1)	7 **G3**	8	3	19
		4	(1)	3	2	0	6
	Total						43
4	Onda 5 pred 3	1	0	0	0	(1)	1
		2	0	0	0	(2)	2
		3	0	0	4	4	8
		4	(1)	(6)	9 **G4**	16	32
	Total						43

Fonte: Elaborada pelos autores (2016).

usado para definir a estabilidade foi o de admitir até um erro de classificação para um dos pares de ondas, para cima no caso grupo G1 ou para baixo no grupo G4. Nos grupos G2 e G3, são admitidos até dois erros de classificação, desde que um seja para cima e outro seja para baixo.

Seguindo este critério, portanto, concluímos existir quatro grupos de classificação estáveis G1, G2, G3 e G4, mostrados na Tabela 8, que indicam quatro grupos de escolas estavelmente classificadas, respectivamente, nos quartis 1, 2, 3 e 4. As não estabilidades, incluindo as instabilidades severas, estão indicadas por círculos. Assim, no caso de Língua Portuguesa, das 172 escolas analisadas, nota-se 138 escolas estavelmente classificadas conforme suas medidas de VA segundo quatro grupos, G1, G2, G3 e G4, de menores para maiores valores de VA. Por outro lado, 34 escolas apresentam não estabilidade de classificação. Resultados similares são encontrados para Matemática, ver Tabela 9.

Os cenários apresentados nas Tabelas 8 e 9 sugerem uma razoável estabilidade de classificação das escolas quando se consideram os modelos de VA produzidos para os três pares de ondas indicados. E, de fato, essa percepção é reforçada pelas correlações de Spearman entre as medidas de VA para as os pares de ondas 5 pred 3 e 3 pred 1, cerca de 0,6 para ambas as disciplinas, observadas para as escolas classificadas nos grupos de G1 a G4 (estáveis), e que correspondem a cerca de 80% das escolas em ambas as disciplinas.

Essa estabilidade parece ser, sob o nosso ponto de vista, razoável para uma decisão que envolva intervenção, por exemplo, e que contemple esses quatro grupos totalmente distintos de escolas – maior estabilidade pode ser alcançada, ainda, restringindo-se as análises às escolas com mais alunos presentes ao teste.

Sobre as demais escolas não estáveis, algumas se encontram em uma zona de indefinição, outras, em uma zona de instabilidade severa, cerca de 13 escolas em ambos os casos, Língua Portuguesa e Matemática, mas não necessariamente as mesmas. Neste último caso, o diagnóstico pode ser que, de fato, haja instabilidade na eficácia dessas escolas entre o ciclo de alfabetização e os dois anos finais do 1º segmento do Ensino Fundamental.

No caso do modelo 4, as correlações de Pearson entre as medidas de VA produzidas em Língua Portuguesa e Matemática mostram variações situadas entre 0,500 e 0,754, dependendo dos diferentes pares de onda. A maior correlação é aquela observada quando se considera o par 5 pred 1. Evidentemente, esses níveis de

correlação, apesar de apreciáveis, são muito inferiores aos que se observam quando se considera apenas medidas de *status*, que variam de 0,930 para a primeira onda até 0,970 para a 5ª onda. Isto é esperado, tendo em vista que os modelos de *Status* que procuram medir a eficácia escolar "descontam" os efeitos do contexto, e os modelos de VA descontam tanto os efeitos de contexto quanto os das condições iniciais dos alunos.

Não foi objeto de estudo aqui, mas seria interessante uma investigação mais detalhada da relação entre o VA medido para Língua Portuguesa e o valor medido para Matemática. É possível, por exemplo, que frente a uma realidade de baixo aprendizado no início de uma etapa escolar, determinadas escolas deliberadamente adotem uma estratégia de privilegiar uma das disciplinas em detrimento da outra.

Tabela 9. Estabilidade da classificação das escolas para o Modelo 3 (VA) segundo os quatro quartis para as ondas 5 pred 3, 3 pred 1 e 5 pred 1 – Matemática.

Quartil VA Onda 5 pred 1			Onda 3 pred 1				Total
			1	2	3	4	
1	Onda 5 pred 3	1	13	10	(3)	(1)	27
		2	7	1	0	0	8
		3	(7)	0	0	0	7
		4	(1)	0	0	0	1
	Total						43
2	Onda 5 pred 3	1	0	1	8	(2)	11
		2	1	6	10	(1)	18
		3	8	5	0	0	13
		4	(1)	0	0	0	1
	Total						43
3	Onda 5 pred 3	1	0	0	0	(2)	2
		2	0	0	8	5	13
		3	0	7	8	0	15
		4	(5)	8	0	0	13
	Total						43
4	Onda 5 pred 3	1	-	0	0	(3)	3
		2	-	0	0	(4)	4
		3	-	0	1	7	8
		4	-	(5)	5	18	28
	Total						43

G1 G2 G3 G4

Fonte: Elaborada pelos autores (2016).

5 Conclusão

A decisão sobre qual modelo de eficácia escolar é o mais apropriado dependerá de restrições técnicas, políticas, e, consequentemente, do desenho do sistema de avaliação. É importante que os gestores tornem explícitos seus objetivos, e que os objetivos de cada modelo também estejam explicitados para os gestores, tendo em vista que escolas avaliadas como as melhores, segundo um determinado modelo, possam não ser com respeito a outros. Assim, as medidas de eficácia devem ser usadas sempre como um elemento a mais para a intervenção nos sistemas escolares, nunca dissociadas dos resultados brutos que podem ser interpretados pedagogicamente. Além disso, outros aspectos do sistema de medição precisam ser considerados. Um deles é o possível viés ocasionado pela não participação de parte significativa dos alunos em quaisquer das avaliações. Essa participação precisa ser monitorada e considerada nos resultados. Isso pode ser feito por meio de modelos ou penalizações nos indicadores.

Quanto à estabilidade, os resultados aqui encontrados apontam na seguinte direção: há estabilidade e informação suficientes para que os modelos de VA possam ser empregados com razoável êxito quando se deseja comparar as escolas segundo sua eficácia, desde que as escolas sejam agrupadas segundo suas medidas de VA. Não é necessário, portanto, ter tantas medidas sucessivas, como no caso do Geres, mas uma medida prévia obtida no início da etapa escolar considerada e uma ao final; talvez, ainda, uma intermediária, caso se deseje avaliar o processo escolar em uma etapa intermediária, e se tenha um projeto claro de intervenção que utilize essa medida intermediária. No entanto, a medida de eficácia para a escola, que considera todo o primeiro segmento do Ensino Fundamental, deveria ser construída a partir da medida inicial e a final somente – pelo menos, para a maioria das escolas. Além disso, os resultados demonstram que uma classificação em níveis, no caso, quatro diferentes níveis ordenados, é muito mais estável e, portanto, racional para ser empregada em uma intervenção do que uma classificação (ranqueamento) geral para todas as escolas. Por outro lado, há de se conhecer melhor as limitações sob as quais as comparações produzidas pelas medidas de VA sejam adequadas. Por exemplo, qual é o nível de heterogeneidade que se admite para o conjunto de escolas comparadas pelas medidas de VA? Será possível, ou desejável, comparar a eficácia da maioria das escolas públicas que atendem alunos de baixa condição socioeconômica com escolas altamente seletivas, como por exemplo, os colégios militares?

A análise comparativa entre os modelos de VA e modelos de *Status* sugere que há diferenças substanciais de classificação das escolas entre os dois modelos, especialmente na disciplina de Matemática. Nesse caso, ao se avaliar a qualidade

da escola apenas por medidas ao final de uma etapa escolar, pode-se estar desconsiderando substancialmente o esforço realizado pela escola vis-à-vis as variáveis de contexto que influem nos resultados finais. É evidente que não sugerimos a simples substituição das medidas de *status* pelas medidas de VA. Acreditamos que a principal informação a ser prestada à sociedade deve ser os níveis de proficiências alcançados pelos alunos ao final das etapas escolares, explicados por meio de uma interpretação pedagógica e, ainda, contextualizados por meio de medidas como as da condição socioeconômica dos alunos e, também, por meio de medidas relevantes de processos escolares. Mas, para amortizar os conflitos que os resultados brutos trazem, e não ser injusto com o esforço da comunidade escolar, nos mais variados contextos sociais e políticos, uma medida direta da qualidade da escola pode ser uma informação útil, senão para a comunidade escolar se espelhar nos resultados, pelo menos para que os gestores possam formular suas políticas públicas adequadamente. Os desafios para a produção de uma medida assim são enormes, mas acreditamos que as medidas de VA podem responder a esse desafio.

Dessa forma, parece-nos viável o emprego dos modelos de VA, pelo menos no primeiro segmento do Ensino Fundamental, pela constituição de boa parte dos sistemas de avaliação no Brasil, pois a maioria já se constitui de uma avaliação ao final do 5º ano do Ensino Fundamental (Prova Brasil e avaliações estaduais, por exemplo), e começa a crescer a avaliação ao final do 3º ano (ANA e outras avaliações estaduais da alfabetização), faltaria apenas uma avaliação inicial, que também poderia ocorrer ao final da Educação Infantil.

É essencial, no entanto, observar que, para se empregar com êxito os modelos de VA, é preciso que haja estratégias de comunicação entre os resultados dos testes e outros instrumentos contextuais, de tal forma que os alunos possam ter suas proficiências e as demais informações relevantes reunidas em uma mesma base de dados, por meio de uma identificação única que permeie todas as avaliações.

O VA calculado para a disciplina de Matemática parece ser uma medida mais fidedigna do esforço da escola do que o calculado para a disciplina de Língua Portuguesa. Esse fato está sintonizado com a ideia de que o aprendizado da Matemática é mais dependente de fatores escolares do que o aprendizado de Língua Portuguesa. Em uma possível combinação das duas medidas, mais peso poderia ser dado para a medida obtida com a avaliação do aprendizado em Matemática, se for do interesse dos gestores, a produção de uma medida global da qualidade da escola. Por outro lado, se o interesse estiver no diagnóstico das escolas, para efeitos de intervenção pedagógica, e não no ranqueamento, as duas medidas devem ser tratadas de forma independente.

Chamamos a atenção para algumas das limitações deste estudo:

Em primeiro lugar, deve-se notar ele se restringe ao primeiro segmento do Ensino Fundamental, principalmente das escolas públicas, que prevalecem na amostra. Além disso, o estudo da estabilidade foi realizado para apenas uma coorte de alunos. Trata-se, portanto, de uma estabilidade interna à coorte avaliada. São necessários estudos que avaliem diferentes coortes ao longo do tempo para que a estabilidade entre as coortes seja avaliada. Esse trabalho está sendo preparado para uma futura publicação. Não exaurimos aqui todos os modelos de VA que podem ser empregados.

Mesmo não sendo objeto de investigação neste estudo, um monitoramento do atrito ou perda de alunos ao longo do período pode ser indicado para correção das medidas de VA quando e se assim se fizer necessário, principalmente para escolas muito seletivas e/ou que apresentem muita perda de alunos ao longo dos anos escolares.

Value-Added Models for the Measurement of the Effectiveness of GERES Schools

Abstract

Using data from the GERES longitudinal research program, the study proposes the comparison of statistical models of varying degrees of complexity to determine the effectiveness of elementary schools. The purpose of the comparison is to determine whether a greater degree of complexity is justified in terms of improved accuracy and if there are differences between the models regarding their consistency and stability when portraying school performance. The two simplest models for the contribution of the school to student proficiency, named status models, use as a proxy for prior proficiency either a measure of the school's average socioeconomic status or a measure of the socioeconomic condition of each student. The other two, called value-added models-VA, incorporate real prior proficiency measures, enabling them to describe the learning gain attributable to the school over the period under study. The results indicate high correlations between the VA models and a low correlation between these and the status models, showing that the gain in precision with the addition of a measure of the socioeconomic condition of each student is small. It is shown that for about 80% of schools the measures of VA across time are stable, suggesting that school effectiveness is a reasonably stable characteristic and that a measure of VA can contribute to the comparison of schools and the definition of interventions, at least during the first years of elementary schooling.

Keywords: *Educational assessment. Value added. Multilevel models.*

Modelos de valor agregado para medir la eficacia de las escuelas Geres

Resumen

A partir de la investigación longitudinal GERES, el estudio propone comparar diferentes modelos estadísticos con grados variados de complejidad para determinar la eficacia de escuelas primarias. El propósito de la comparación es determinar si un mayor grado de complejidad se justifica en términos de una mayor precisión, y si hay diferencias entre los modelos en su consistencia y capacidad para retratar de manera estable el rendimiento de la escuela. Los dos modelos más simples de la contribución de la escuela para la proficiencia del estudiante, llamados modelos de status, incorporan una medida del nivel socioeconómico promedio de la escuela o una medida de la situación socioeconómica de cada alumno como proxy de la proficiencia previa. Los otros dos, llamados Modelos de Valor Agregado - VA, incorporan medidas de proficiencia prévia, que los hacen modelos capaces de describir la ganancia de aprendizaje atribuible a la escuela durante el período estudiado. El estudio indica una alta correlación entre los modelos de VA, ysu baja correlación con los modelos de status, lo que demuestra que la ganancia de precisión con la adición de una medida de la situación socioeconómica de cada alumno es pequeña. Resulta que alrededor del 80% de las escuelas se mantuvieron estables para las diferentes medidas temporales de VA, lo que sugiere que la eficacia es, en realidad, una característica razonablemente estable en el tiempo y que el VA puede ayudar en la comparación de las escuelas y en la definición de las intervenciones, al menos en el primer segmento de la escuela primaria.

Palabras clave: *Evaluación educativa. Valor agregado. Modelo multinivel.*

Referências

AMREIN-BEARDSLEY, A. *Rethinking value-added models in education*: critical perspectives on tests and assessment-based accountability. New York: Routledge, 2014.

BALLOU, D. Value-added assessment: lessons from Tennessee. In: Lissitz, R. W. (Ed.). *Value-added models in education:* theory and application. Maple Grove, MN: JAM, 2005. p. 272-97.

BRAUN, H.; CHUDOWSKY, N.; KOENIG, J. (Eds.). *Getting value out of value-added: report of a workshop* Washington, DC: National Academies Press, 2010.

BRAUN, H.; WAINER, H. Value-added model. In: RAO, C. R.; SINHA RAY, S. (Eds.). *Psychometrics*. Amsterdam: Elsevier, 2007. p. 867-92. (Handbook of statistics, 26).

BROOKE, N.; BONAMINO, A. M. C. (Orgs.). *Geres 2005:* razões e resultados de uma pesquisa longitudinal em eficácia escolar. Rio de Janeiro: Wallprint, 2011.

BUCHMANN, C. Measuring family background in international studies of education: conceptual issues and methodological challenges. In: GAMORAN, A.; PORTER, A.C. (Eds.). *Methodological advances in cross-national surveys of educational achievement*. Washington, DC: National Academy Press, 2002. p. 150-97.

FELDMAN, B.; RABE-HESKETH. Modeling achievement trajectories when attrition is informative. *Journal of Educational and Behavioral Statistics*, v. 37, n. 6, p. 703-36, Dec. 2012. doi:10.3102/1076998612458701

FERRÃO, M. E. School effectiveness research findings in the Portuguese speaking countries: Brazil and Portugal. *Educational Research Policy and Practice*. v. 13, n. 1, p. 3-24, Feb. 2014. doi:10.1007/s10671-013-9151-7

FERRÃO, M. E.; COUTO, A. Indicador de valor acrescentado, tópicos sobre consistência e estabilidade: uma aplicação ao Brasil. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 21, n. 78, p. 131-64, jan./mar. 2013.

_____. The use of a school value-added model for educational improvement: a case study from the Portuguese primary education system. *School Effectiveness and School Improvement*, v. 25, n. 1, p. 174-90, 2014. doi :10.1080/09243453.2013.785436

GRAY, J. et al. A multi-level analysis of school improvement: changes in schools' performance over time. *School Effectiveness and School Improvement*, v. 6, n. 2, p. 97-114, 1995.

LECKIE, G.; GOLDSTEIN, H. The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society. Series A: Statistis in Society*, v. 172, n. 4, p. 835-51, Oct. 2009.

RAUDENBUSH, S. W.; BRYK, A. *Hierarquical linear models*. 2. ed. London: Sage, 2002.

REARDON, S.; RAUDENBUSH, S. Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, v. 4, n. 4, p. 492-519, 2009. doi:10.1162/edfp.2009.4.4.492

RUBIN, D. B.; STUART, E. A.; ZANUTTO, E. L. A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral*, v. 29, n. 1, p. 103-16, 2004.

SOARES, J.F.; XAVIER F. P. Pressupostos educacionais e estatísticos do IDEB. *Educação e Sociedade*, v. 34, n. 124, p. 903-23, jul./set. 2013. doi:10.1590/S0101-73302013000300013

SOARES, T. M. Evaluating the stability of the value-added scores obtained through a longitudinal multilevel IRT model. In: WORLD STATISTICS CONGRESS OF THE INTERNATIONAL STATISTICAL INSTITUTE, 60., 2015. Rio de Janeiro. *Anais...* The Hague, The Netherlands: InternationalSta tisticalInstitute, 2015. Disponível em: <http://www.isi2015.org/components/com_users/views/registration/tmpl/media/uploadedFiles/abstracts/149/778/IPS076-P1-A.pdf>. Acesso em: 15 jan. 2016.

TEKWE, C. D. et al. An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics,* v. 29, n. 1, p. 11-36, Mar. 2004. doi:10.3102/10769986029001011

Informações dos autores

Tufi Machado Soares: Doutor em Teoria Matemática de Controle e Estatística pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Professor Titular do Instituto de Ciências Exatas e do Programa de Pós-Graduação em Educação da Universidade Federal de Juiz de Fora – UFJF. Contato: tufi@caed.ufjf.br

Alicia Bonamino: Doutora em Educação pela Pontifícia Universidade Católica do Rio de Janeiro. Professora Associada do Departamento de Educação da PUC-Rio. Contato: alicia@puc-rio.br

Nigel Brooke: Doutor em Estudos do Desenvolvimento pelo Institute of Development Studies da University of London. Professor Convidado da FAE/UFMG e Pesquisador do Grupo de Avaliação e Medidas Educacionais – GAME/UFMG. Contato: n.brooke@terra.com.br

Neimar da Silva Fernandes: Graduado em Ciências Exatas pela Universidade Federal de Juiz de Fora – UFJF. Auxiliar de Pesquisa no Centro de Avaliação de Políticas Públicas da Educação – CAEd/UFJF. Contato: neimar@caed.ufjf.br