Relatório da Avaliação do Processo de Inovações no Ciclo Básico e seu Impacto Sobre a Situação do Ensino-Aprendizagem na Região Metropolitana de São Paulo: Comentário

Marlaine Lockheed*

Agradeço o convite que me foi feito para comentar essa avaliação sistemática, pouco comum e complexa, do impacto da reforma do ensino no nível de aprendizagem das crianças do primeiro grau. Os pesquisadores do estudo merecem congratulações por terem implantado a avaliação do impacto e, principalmente, por terem escolhido manter seu projeto original de acompanhar as crianças durante três anos. No mundo inteiro, avaliações longitudinais são raras e as informações levantadas nesta avaliação são certamente preciosas para a pesquisa que servirá de base para o enriquecimento das análises futuras.

De início, devo declarar que não li os relatórios detalhados em português nos quais a versão em inglês se fundamenta. Percebo que a versão em inglês é, sem dúvida, apenas um resumo sumário de estudos muito bem documentados. Desculpo-me por desconhecer o português, espero que meus comentários, inteiramente baseados no resumo em inglês, tenham alguma utilidade para a série completa de estudos. Isto posto, começarei daqui.

* Departamento de Desenvolvimento Humano do Banco Mundial. Texto apresentado no Seminário organizado pela Secretaria de Estado da Educação, São Paulo, 26-28 de fevereiro, 1996.

Qualquer comentário a respeito de uma avaliação de impacto levantará pelo menos quatro questões principais, e organizarei meus comentários em torno delas:

- Qual o objetivo desta avaliação?
- O perfil desta avaliação foi apropriado para o objetivo proposto? (e uma subquestão, o processo de amostragem foi adequado?)
- A avaliação foi implantada de forma apropriada?
- Os dados foram analisados de forma adequada?

1. Qual o objetivo da avaliação?

Avaliações de impacto sobre o programa procuram estabelecer a causalidade entre a participação no programa e as mudanças em alguns pontos relevantes. Estabelecer que o programa teve um impacto no nível de aprendizagem dos alunos significa estabelecer que o programa produz efeitos que são diferentes daqueles que teriam ocorrido sem o programa ou com um programa alternativo.

No caso da avaliação do impacto de inovações no Ciclo Básico, o objetivo era o de estabelecer uma relação de causa entre a participação de um estudante num dos três tipos de programas escolares e seu crescimento no nível de aprendizagem. Os três programas eram: (a) o Ciclo Básico (iniciado em 1983), (b) o Ciclo Básico e a Jornada Única (iniciado em 1988), que abrangeu cerca de 75% das escolas em 1991, e (c) o Programa de Escola-Padrão (iniciado em 1991) que abrangeu cerca de 10% das escolas da região metropolitana de São Paulo que implementaram a Jornada Única. A implantação do Programa da Escola-Padrão ocorreu em duas etapas, 1992 e 1993.

Há dois pré-requisitos para avaliar o impacto de um programa: o de que os objetivos do programa são suficientemente bem articulados para tornar possível a especificação de medidas de realização das metas, e o de que o programa foi suficientemente bem implementado para que não exista dúvida que seus componentes de análise foram dirigidos às metas certas, neste caso, os alunos.

Os objetivos dos três programas foram bem articulados: elevar o nível de aprendizagem dos alunos, reduzir a evasão e a repetência. Estes são objetivos claros e preenchem o primeiro requisito de uma avaliação de programa. O resumo da avaliação contém respostas às perguntas seguintes:

- *Qual foi o impacto da participação no programa no nível de aprendizagem dos alunos?*

Especificamente:

• Os programas de Jornada Única e Escola Padrão melhoram a aprendizagem das crianças nos três primeiros anos da escola?

• O desempenho dos alunos em Matemática e Português nas escolas que possuem o Ciclo Básico e a Jornada Única é mais alto do que o dos alunos do Ciclo Básico?

• O desempenho dos alunos em Matemática e Português nas Escolas-Padrão é mais elevado do que o dos alunos das escolas com Ciclo Básico ou com Ciclo Básico com Jornada Única?

• O desempenho dos alunos em Matemática e Português nas Escolas-Padrão (92) é mais elevado do que o dos alunos das escolas com Ciclo Básico, com Ciclo Básico com Jornada Única e Escola-Padrão (93)?

O resumo não dá informações sobre o impacto dos três programas no tocante à evasão ou à repetência.

O segundo requisito é que os programas foram implantados. De acordo com o estudo, os três programas de ensino – Ciclo Básico, Jornada Única e Escola-Padrão – foram implantados em vários níveis nas escolas. Devido a esta falta de uniformidade na implantação do programa, várias etapas serão necessárias para medir qual parte do programa foi implantada. Além disso, no caso da Escola-Padrão, o programa foi implantado em algumas escolas apenas em 1992 e em muitas outras em 1993; essa variação foi considerada e um quarto "estrato" foi identificado para a análise.

Logo uma segunda série de questões foi levantada:

Quais os elementos específicos do programa que resultaram num rendimento maior na aprendizagem?

Especificamente:

• Quais fatores nível-estudante "previam" rendimento?
• Quais fatores nível-escola "previam"rendimento?

Logo, no que diz respeito ao objetivo da avaliação, é claro e viável.

2. O PERFIL DA AVALIAÇÃO É APROPRIADO PARA CUMPRIR ESTE OBJETIVO?

A segunda questão refere-se ao perfil da avaliação. Ele é capaz de discernir o impacto? As estratégias de avaliação são diferentes, por necessidade, para programas que abrangem a população toda e os que afetam apenas parte dela. Para o primeiro, pensa-se no chamado *antes* e *depois* do estudo, ou examinar as tendências ao longo do tempo, uma vez ou

repetidamente, para determinar o impacto do programa. De forma alternativa, se o programa abrange toda a população, mas não é uniforme, deve-se poder usar esta falta de uniformidade para analisar seu impacto, usando técnicas de corte transversal. Existem outras opções que podem ser utilizadas quando um programa abrange apenas parte da população ou se é introduzido ao longo do tempo. Estas opções variam de controle aleatório ou "verdadeiros" perfis experimentais a perfis quase-experimentais que utilizam controles estatísticos ou elaborados (Rossi and Freeman 1989).

A avaliação de impacto de São Paulo não utiliza o planejamento de controle aleatório. Os alunos não são colocados num programa específico através de um procedimento aleatório. Nem as escolas foram escolhidas aleatoriamente para fazer parte dos programas de Ciclo Básico, Jornada Única ou Escola-Padrão. As escolas decidiram participar ou não, e os alunos (ou seus pais) escolheram freqüentar uma escola ou outra.

O mais próximo a que se pode chegar de um perfil de controle aleatório é elaborar grupos de controle que são equivalentes aos grupos experimentais em todos os aspectos relevantes. Esta abordagem de perfil quase-experimental significa comparar os participantes do projeto com pessoas que são "similares" em termos de idade, sexo, educação, renda, região, *status* social e qualquer outro tipo de características que possa ser relevante para determinar o impacto da intervenção. Comparações são então feitas entre o grupo experimental e o grupo de controle elaborado. Isto não foi feito na avaliação em São Paulo.

Uma terceira estratégia é a de utilizar controles estatísticos para as diferenças das populações. Esse perfil foi utilizado na avaliação de São Paulo. Controles estatísticos foram introduzidos para diferenças entre escolas em termos de idade, sexo, cor, freqüência em pré-escola dos alunos e classe na qual estavam matriculados. A avaliação, porém, não considera influência da seleção. Já que os alunos não foram aleatoriamente colocados em programas escolares, eles ou seus pais podem ter escolhido um programa específico. O controle estatístico para a seletividade pode ser introduzido na análise, mas isto não foi feito na avaliação de São Paulo.

Recomendação: Futuramente, as análises deveriam testar os efeitos da seleção. Uma metodologia convencional seria a da abordagem em duas etapas de Heckman (Heckman,1979).

2a – Será possível generalizar a partir da amostragem para a população de alunos em cada programa?

Aqui, a questão é: a amostra dos alunos estudados na avaliação representa de forma adequada cada um dos programas? A resposta a esta

pergunta será "sim" se o procedimento para selecionar uma amostra aleatória de participantes de cada um dos três programas foi utilizado. A versão em inglês não descreve o processo pelo qual 60 escolas ou 2 classes de CB1 de cada escola foram escolhidas; o estudo observa que todas as crianças das 120 classes foram escolhidas. Tendo em vista o cuidado com o qual foi feita esta avaliação, imaginei que um procedimento científico de amostragem tenha sido empregado. Isto teria envolvido tipicamente (a) a designação dos programas (*strata*) **ex ante,** (b) a enumeração das escolas em cada programa (*strata*), (c) seleção aleatória de escolas em cada estrato, (d) seleção aleatória de classes em cada escola, e (e) seleção aletória de estudantes em cada classe. Como os procedimentos de seleção fortuita não foram utilizados, os resultados não poderão ser totalmente generalizados.

A versão em inglês faz menção de uma estratégia de amostragem que poderia afetar os resultados. Isto é, a amostragem que se fundamenta numa variável dependente. Especificamente, um procedimento de amostragem estratificada foi utilizado, pelo qual além do estrato do "programa", as escolas foram estratificadas de acordo com o volume de matrículas, transferências, níveis de aprovação, repetência, evasão e situação geográfica. Logo, de acordo com o relatório, as escolas foram selecionadas com base em duas variáveis possíveis: evasão e repetência. Este fato pode comprometer os resultados, a menos que técnicas analíticas apropriadas sejam utilizadas.

Recomendação: Futuramente, as análises deveriam testar a probabilidade de evasão de vários programas, utilizando uma análise criteriosa de variáveis múltiplas (Bishop, Fienberg e Holland, 1975).

3. A AVALIAÇÃO FOI IMPLANTADA DE FORMA APROPRIADA?

Para que qualquer avaliação tenha significado, os dados analisados devem ser tratados com o cuidado necessário para que sejam confiáveis e válidos. E, numa avaliação longitudinal, todo esforço deve ser feito para garantir que as diferenças entre os programas no tocante a seu impacto aparente não possam ser atribuídas a diferentes níveis de retenção entre eles.

Qualidade de dados. O relatório resumido em inglês fornece informações suficientes e sugere que a qualidade dos dados é boa. Vários índices apropriados foram desenvolvidos para medir o impacto dos programas e controlar as diferenças entre os alunos da amostra nos quatro estratos. Mais informações a respeito da confiabilidade dos testes poderia ter sido fornecida, mas isto se encontra sem dúvida em relatórios mais completos.

Amostra de retenção. O elemento que mais influi numa avaliação longitudinal é que os programas que são avaliados têm diferentes potenciais. Isto é, um programa pode ser melhor para manter as crianças na escola do que outro. De acordo com o relatório, as escolas foram escolhidas com base em, *inter alia,* evasão e repetência. Os dados apresentados no relatório, e que me foram fornecidos a pedido, mostram que aproximadamente 40 por cento da amostra original de alunos abandonaram a escola ao longo dos três anos. A taxa mais alta de retenção foi de 61,7 por cento para alunos do Ciclo Básico (estrato 4) comparado com 59,7 por cento dos alunos da Jornada Única (estrato 3) e aproximadamente 58,7 por cento dos alunos da Escola-Padrão (estratos 1 e 2). Mas estas diferenças na retenção da amostra entre os programas são mínimas, e podem simplesmente refletir os critérios para a seleção das escolas no início. O fato de que o percentual de alunos que abandonaram os vários programas seja semelhante não significa que os programas não tenham impacto na evasão ou que este impacto não afete os resultados.

Recomendação: As análises precisam testar se as evasões de um programa diferem de forma significativa das evasões de um programa diferente. Se um bom aluno abandona um programa, enquanto alunos de menor desempenho abandonam outro programa, o impacto aparente dos dois programas poderia ser bem diferente. A avaliação precisará comparar as características de lançamento de evasões entre os quatro tipos de programa para verificar se existem diferenças.

4. Os dados foram analisados de forma adequada?

O método de análise para uma avaliação de programa deveria ser escolhido de acordo com dois critérios: o primeiro, o de permitir que a avaliação responda às perguntas colocadas no ponto de partida e o segundo, o de se adequar aos dados. Neste caso, a questão colocada diz respeito à causalidade entre a participação num programa particular e o crescimento no nível de aprendizagem, e os dados estão organizados de forma hierárquica. A versão em inglês da análise não fornece informações suficientes nos modelos explícitos que foram testados para que se possa afirmar se eram adequados ou não. Todavia, as observações seguintes têm o intuíto de estimular a discussão nestes pontos.

Comparação entre dados e análise

Dados organizados hierarquicamente. O método apropriado para analisar estes dados é a técnica de modelo linear hieráquico que parece ter

sido utilizado aqui. Este método é apropriado devido à série de dados hierarquicamente desigual de alunos nas classes, nas escolas, nos programas. Uma análise em três fatores seria mais apropriada.

Seletividade da amostra. Além disso, a análise precisaria fazer correções no tocante à seletividade da amostra e à probabilidade de evasão antes de avaliar o impacto do programa no desempenho. Infelizmente, a análise não tratou deste aspecto. No entanto, os pesquisadores enfrentaram um dilema. As correções para a seleção da amostra e a evasão podem ser feitas a partir de modelos de nível único, já que nenhum algoritmo multinível, incorporando esses procedimentos de correção, está disponível atualmente, ao meu conhecimento. Mas modelos de nível único não avaliam corretamente erros para dados organizados hierarquicamente de forma desigual, como nesta série de dados, e devem portanto ser evitados.

Recomendação: Teste para seletividade da amostra, mesmo se for a de nível único (cf. Limdep), deve ser utilizado.

Comparação entre questões de avaliação e análise

As questões da avaliação centraram-se no impacto dos programas no crescimento do nível de aprendizagem ao longo dos três primeiros anos da escola primária, a eficácia da comparação dos vários programas, e os fatores relevantes para estes efeitos.

Impacto dos programas no nível de aprendizagem. A primeira questão da avaliação era a respeito do impacto dos programas no nível de aprendizagem ao longo dos três primeiros anos. Uma implicação desta questão é a variável dependente – medir o nível de aprendizagem – que precisaria ser elaborada de forma a indicar crescimento. Isto aparentemente foi feito utilizando técnicas de Teoria de Resposta ao Item.

O capítulo de resultados da versão em inglês mostra que o nível de aproveitamento médio em Português e Matemática permanece relativamente estável no âmbito de cada tipo de escola ao longo dos três anos de avaliação. Isto não parece plausível, a menos que a intenção seja demonstrar que não existe aprendizagem na escola.

Recomendação: Os escores da escala TRI devem ser incluídos neste capítulo.

Comparação dos programas no tocante a nível e crescimento

A segunda série de questões abrange o impacto comparativo dos programas. O relatório mostra que nível de escores para Português e Matemática é mais elevado para alunos das Escolas-Padrão (Estratos 1 e 2),

em seguida para estudos do Ciclo Básico e Jornada Única (Estrato 3) e mais baixos para alunos das escolas de Ciclo Básico (estrato 4). Em Matemática, os alunos das Escolas-Padrão, cujo programa foi implantado em 1992, têm escores muito mais altos em Matemática do que as Escolas-Padrão com programas implantados em 1993. Não parece haver escores de TRI para isto, logo a comparação só pode ser feita por corte transversal.

A questão crucial para a avaliação do impacto do programa é a de que – sem alterar outros dados – os três programas têm um impacto diferenciado no crescimento do desempenho escolar. Esta questão é tratada superficialmente na versão em inglês. Já que os modelos completos não estão detalhados, fica muito difícil entender o que realmente representam os quadros (Tabelas 6 e 8).

Recomendação: Os modelos completos que são testados devem ser detalhados para permitir que o leitor entenda os resultados.

Análise complementar

Além das análises que foram elaboradas para responder às perguntas da avaliação, o relatório inclui também um corte que usa uma técnica de modelo linear hierárquico de dois fatores (aluno e escola) para separar a variância no desempenho do aluno. Não fica claro qual é o propósito da análise ou exatamente o que foi feito. Uma abordagem típica para elaborar este tipo de análise deve separar a variância no desempenho dentro do componente "entre escola" e o componente "entre aluno", usando o que é chamado normalmente de modelo"nulo" ou "vazio". A avaliação parece fazer esta análise e mostra que menos de 20 por cento da variância no desempenho em Português e 30 por cento de variância em Matemática se deve a fatores "entre escola". Esta proporção cai ao longo dos três anos para 11 por cento de variância em Português e 13 por cento de variância em Matemática. Dado que é um modelo de dois fatores, o resto da variância pode ser atribuído a fatores "entre alunos". Porém, o relatório não fornece indicações se ele apresenta os resultados do modelo "nulo" ou "vazio" ou resultados de outros modelos que incluem controles para características de fator aluno e tipo de programa.

Recomendação: O estudo deve vincular de forma mais estreita esta análise às questões da avaliação e descrever os modelos de forma mais completa.

Explicação dos efeitos do programa

O terceiro objetivo da avaliação de impacto era o de identificar as características dos programas que explicam as diferenças. Como foi relatado no resumo, a análise parece introduzir controles para as características do

aluno simultaneamente com variáveis explicativas do fator escola. Além disso, a Etapa 2 parece mudar do modelo fator-2 para o modelo fator-3, ou alunos nas escolas dentro dos estratos. Isto não fica claro no resumo em inglês.

Os resultados substantivos das análises de regressão são muito interessantes: o compromisso do corpo docente, a liderança do diretor, a estabilidade do professor no processo ensino-aprendizagem aparecem todos determinantes importantes do sucesso. O relatório não fornece indicações se estes aspectos são mais característicos de um programa ou de outro.

Recomendação: As análises para explorar as razões para efeitos do programa precisarão vincular variáveis "escolas eficientes" a programas específicos aos quais estão ligadas para atender a questão da avaliação.

5. Conclusão

Esta avaliação do impacto terá de ser considerada um estudo comparativo de regressão entre os quatro tipos de escolas. Tirará proveito dos comentários a respeito das análises que são discutidas no último capítulo do estudo e fazendo os acertos necessários para a seletividade do programa e taxas de retenção de programas diferenciados. Sem estes acertos, será difícil determinar se os tipos de escolas têm um impacto diferente no desempenho. Ao mesmo tempo, o aspecto longitudinal do estudo permite a identificação de algumas características importantes referentes ao crescimento do desempenho. Com análises adequadas, a avaliação do impacto poderá fornecer dados com relação aos fatores das escolas que reduzirão as diferenças de gênero e cor no desempenho.

Referências

BISHOP, Y.M.M; S. E. FIENBERG e P. HOLLAND. 1975. **Discrete Multivariate Analysis.** Cambridge, Mass: MIT Press.

HECKMAN, JAMES, 1979. "Sample Selection Bias as Specification Error". **Econometrica** 47 (Janeiro): 153-61

ROSSI, PETER e HOWARD FREEMAN, 1989. **Evaluation: A Systematic Approach.** Newbury Park, Calif.: Sage Publications.