# Avaliação de impacto no Brasil: é possível mensurar impactos de programas de formação docente?

ADRIANA BAUER*

## RESUMO

O presente artigo visa a compartilhar as vicissitudes de uma proposta de avaliação de impacto de um programa educacional, objeto atual de análise da autora. Tal estudo, provisoriamente intitulado "Avaliação de possíveis impactos de programas de formação docente: a experiência do Programa Letra e Vida", originou-se da preocupação com a escassez de trabalhos que buscassem mensurar impactos de programas educacionais no Brasil. O texto discute o conceito de "avaliação de impacto" e analisa as dificuldades metodológicas inerentes a esse tipo de avaliação. Traz exemplos dos desafios subjacentes à mensuração de impacto em educação que frustraram o projeto inicial de doutoramento da autora e as tentativas realizadas no sentido de superar tais obstáculos. A motivação para a produção do artigo não é, portanto, compartilhar resultados de pesquisa, mas sim propor a discussão das dificuldades encontradas e dos caminhos metodológicos que podem ser utilizados para o estudo de efeitos de programas, visando ao aprofundamento das reflexões sobre a temática.

**Palavras-chave**: Avaliação do programa, Formação de professores, Saresp, Programa Letra e Vida.

* Doutoranda em Educação, Programa de Pós-Graduação da Faculdade de Educação da Universidade de São Paulo e Pesquisadora da Fundação Carlos Chagas (dri_bauer@yahoo.com.br).

## RESUMEN

El presente artículo tiene como objetivo compartir las vicisitudes de una propuesta de evaluación de los impactos de un programa educacional. Este es el objeto actual de análisis de la autora. Tal estudio, con el título provisorio de "Evaluación de los posibles impactos de programas de formación docente: la experiencia del Programa *Letra e Vida*", surgió de la preocupación por la escasez de trabajos que midiesen los impactos de programas educacionales en Brasil. El texto discute el concepto de "evaluación de impacto" y analiza las dificultades metodológicas inherentes a este tipo de evaluación. Aporta ejemplos de los desafíos subyacentes a la medición del impacto en educación, que frustraron el proyecto inicial de doctorado de la autora, y los intentos realizados con el fin de superar tales obstáculos. La motivación para producir el artículo no es, por lo tanto, compartir los resultados de una investigación, sino proponer la discusión de las dificultades encontradas y de los caminos metodológicos que se pueden utilizar para el estudio de efectos de programas, con el objetivo de profundizar las reflexiones sobre este tema.

**Palabras clave**: Evaluación del programa, Formación de profesores, Saresp, Programa *Letra e Vida*.

## ABSTRACT

This article aims at sharing the hardships of an evaluation proposal of the impact of an educational program, the present object of the author's analysis. This study, provisionally named "Evaluation of possible impacts of teacher education programs: the experience of the *Letra e Vida* Program", arose from a concern with the lack of studies that attempted to measure impacts of educational programs in Brazil. The text discusses the concept of "impact evaluation" and analyzes the methodological difficulties inherent to this type of evaluation. It also provides examples of challenges underlying the measurement of impact in education which frustrated the author's initial doctoral project, and the attempts made to try to overcome such obstacles. The motivation for this article, then, is not to share research results, but to propose the discussion of the difficulties encountered and of the methodological approaches that can be used to study the effects of the programs, so as to widen the reflections on this topic.

**Keywords**: Program evaluation, Teacher education, Saresp, *Letra e Vida* Program.

## INTRODUÇÃO

Algumas políticas educacionais implementadas no Brasil, desde meados da década de 1990, destinaram parte dos recursos disponíveis à formação e ao desenvolvimento dos professores, visando, entre outros objetivos, à melhoria da qualidade do ensino.

Como exemplo, pode-se citar o Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério (Fundef) e o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (Fundeb). Primeira iniciativa de política de fundos, estabelecida após o período da redemocratização no Brasil, o Fundef subvinculou 60% dos recursos destinados ao ensino fundamental à formação e ao desenvolvimento dos professores, enquanto o Fundeb manteve a mesma subvinculação de recursos para a formação, ampliando-a a todos os professores de educação básica.

Quer seja para o ensino fundamental, quer contemple toda a educação básica, ambas as iniciativas expandiram consideravelmente a possibilidade dos sistemas educacionais designarem recursos para o desenvolvimento profissional dos professores. Além disso, diversas secretarias estaduais de educação vêm sistematicamente investindo na formação dos professores, visando a mudanças na qualidade do ensino ofertado.

No caso específico do Estado de São Paulo, diferentes ações de formação contínua vêm sendo planejadas e ofertadas, sendo a formação docente apresentada como um dos eixos da segunda gestão do governo Alckmin (2003-2006), inserida em um projeto político mais amplo, focado na "inclusão social e melhoria do ensino" (São Paulo, 2003, p. 13).[1]

Nesse sentido, a qualificação docente e a melhor atuação do professor têm sido considerados fatores preponderantes para o desenvolvimento da qualidade de ensino e do desempenho dos estudantes, por meio do "aprimoramento das práticas escolares" (São Paulo, 2003, p. 15).

Inclusive ao referenciar os sistemas de avaliação como importantes formas de acompanhamento desse "aprimoramento", o documento que lançou as diretrizes para a política educacional do governo Alckmin fez alguns anúncios sobre a necessidade de avaliação das ações implementadas. Não foram especificados,

---

[1] A existência de programas de educação continuada foi mantida no governo José Serra (2007-2010) embora aparentemente com menor ênfase. O apoio a essas atividades faz parte das dez metas para a educação, lançadas pela Secretária de Educação Maria Helena Guimarães de Castro, a serem atingidas até 2010.

no entanto, critérios que pudessem balizar essa avaliação e o monitoramento dos programas implementados.

Apesar de as intenções de acompanhamento e avaliação dos programas estarem divulgadas nos documentos oficiais da Secretaria da Educação do Estado de São Paulo daquela gestão, não foram encontradas referências a essa avaliação em outros documentos pesquisados. Mesmo nos documentos específicos do programa de formação continuada Teia do Saber, que deu origem aos programas de formação da Secretaria da Educação (entre os quais o "Letra e Vida"), não foram encontradas informações que permitissem vislumbrar como se daria sua avaliação.

O fato de tal avaliação não estar especificada nos documentos chamou a atenção da pesquisadora, dada a ênfase que vem sendo colocada pelas políticas de diversas áreas na avaliação dos resultados dos programas como fator essencial para seu redimensionamento.

Para além da análise documental, durante a pesquisa empírica, realizada pela autora no processo de elaboração da sua dissertação (Bauer, 2006), foi possível verificar que são raros os momentos em que programas da Secretaria da Educação do Estado de São Paulo são avaliados formalmente. Além disso, as avaliações, quando realizadas, tendem a focar a implementação do programa ou seus resultados imediatos, deixando de lado o estudo sobre o impacto dos mesmos, entendido aqui como efeitos que se mantêm a longo prazo.

Não seria de esperar que no momento em que o discurso da qualidade, da eficiência e eficácia de programas fosse utilizado como justificativa das ações desencadeadas e a avaliação dos impactos dessas ações fosse incorporada aos desenhos das políticas?

Tal questionamento deu origem a outros: seria possível saber em que medida os esforços realizados, no sentido de melhorar a qualidade da formação docente, surtem efeito sobre a "qualidade do trabalho docente" ou sobre a atuação dos professores? Como avaliar a influência que programas de formação têm acerca da atuação docente e do desempenho dos alunos? Ou, de forma geral, quais seriam os possíveis impactos das ações formativas na prática docente e nos resultados dos alunos?

Observa-se que nos documentos do "Letra e Vida", assim como nos outros pesquisados, foram feitas afirmações a respeito da necessidade de seu acompanhamento e avaliação, mas não são encontradas especificações acerca de mecanismos de avaliação dos resultados e impactos do programa.

Foi a partir dessas questões e da lacuna de propostas de avaliações oficiais do programa que o objeto e problema do estudo proposto pela autora, em nível de

doutoramento, foram definidos: análise de eventuais impactos do Programa Letra e Vida, implantado em 2003 pela Secretaria da Educação do Estado de São Paulo. Para esse estudo a análise focaliza os professores da 1ª série do ciclo de alfabetização (ensino fundamental 1).

Este é um programa de formação para professores alfabetizadores, destinado aos docentes do primeiro segmento do ensino fundamental, mais especificamente aos professores de 1ª e 2ª séries.

O objetivo geral da pesquisa é analisar eventuais impactos do Programa Letra e Vida na atuação dos professores e no desempenho dos alunos. Os objetivos específicos são:

- Identificar evidências, na organização do trabalho em sala de aula, da utilização de princípios metodológicos do "Letra e Vida".
- Analisar eventuais diferenças, no desempenho de alunos de escolas estaduais no Saresp, que possam ser atribuídas ao "Letra e Vida".
- Procurar indícios, nos discursos dos professores, que evidenciem a influência (ou não) da formação recebida em sua prática cotidiana.
- Elencar dificuldades e desafios que se impõem aos professores para fazer uso dos conhecimentos teóricos adquiridos durante o curso em sua prática cotidiana.

Feitos os primeiros delineamentos, a autora passou à análise documental e ao levantamento de dados para realização da pesquisa. Foi quando as primeiras dificuldades e desafios se impuseram, levando-a a refletir sobre a metodologia inerente a esse tipo de avaliação e, ainda mais, sobre o significado conceitual de avaliação de impacto, como será visto a seguir.

## O QUE É AVALIAÇÃO DE IMPACTO?

A avaliação de um programa social pode envolver diversas etapas: **análise da proposta** (examina se o programa é importante e relevante para o objetivo pré-definido e se o desenho está adequado, projeta possíveis resultados, etc.), **da implementação** (avalia se o projeto está sendo conduzido conforme o planejado), **dos resultados** (analisa se o programa implementado atingiu os objetivos previamente definidos) e **dos impactos**, entendidos aqui como resultados e efeitos da intervenção a longo termo e que se mantêm mesmo após o término da intervenção.

Contudo, na vasta literatura existente sobre avaliação de programas, nem sempre essas são as etapas mencionadas e, tampouco, os conceitos utilizados por diversos

autores se equivalem sendo que, muitas vezes, a ideia de impacto está incorporada na avaliação de resultados, e os termos utilizados como sinônimos.

Mesmo dentre os autores que fazem distinção entre "resultados" e "impactos", observa-se que as definições de avaliação de impacto são diversas, havendo pouco consenso, nas referências pesquisadas, sobre o significado do termo.

Michael Scriven, por exemplo, no clássico *Evaluation Thesaurus*, define[2] avaliação de impacto como "uma avaliação focada nos resultados ou retornos do investimento, em vez de no processo, na entrega, ou na avaliação da implementação"[3] (1991, p. 190).

Nesse exemplo, nota-se que a definição de impacto relaciona-se ao foco da avaliação, e pode-se inferir que impactos e resultados (*outcomes*) são indistintamente entendidos pelo autor como "efeitos", ou seja, possuem uma natureza relacional com a intervenção, podendo ocorrer "durante", "ao final" da intervenção ou "posteriormente" (Scriven, 1991, p. 250). O exemplo do autor ilustra a tendência do uso intercambiável entre esses termos, encontrados em parte da literatura destinada à avaliação de programas (Weiss, 1998; Stufflebeam; Webster, 1980), como será exemplificado a seguir.

Mohr (1992), por exemplo, utiliza "análise de impacto" e aponta que impactos ocorrem quando uma intervenção afeta o estado de um objeto ou fenômeno "mais de uma vez":

> Vamos tomar o termo análise de impacto para significar a determinação da extensão em que um conjunto de atividades humanas dirigidas (X) afeta o estado de alguns objetos ou fenômenos ($Y_1$, ..., o $Y_k$) – pelo menos algumas vezes – determinando por que razão os efeitos foram tão pequenos, ou grandes, como acabaram por ser. (p. 1)[4]

---

[2] Como a maioria das citações utilizadas neste artigo provém da literatura estrangeira, optou-se por colocar as citações originais, a fim de preservar a fidedignidade do texto, que poderia ser prejudicada por traduções equivocadas. Contudo, para garantir o acesso à informação a todos os leitores, uma tradução livre foi elaborada pela autora, sempre que recorreu a passagens para reforçar ou exemplificar os argumentos.

[3] "An evaluation focused on outcomes or payoff rather than process, delivery, or implementation evaluation" (Scriven, 1991, p. 190).

[4] "Let us take the term *impact analysis* to mean determining the extent to which one set of directed human activities (X) affected the state of some objects or phenomena ($Y_1$, ..., $Y_k$) – at least sometimes – determining why the effects were as small or large as they turned out to be." (Mohr, 1992, p. 1)

Para esse autor, isso significa que para poder atribuir um efeito (o estado de algum objeto ou fenômeno) a uma determinada causa (atividade humana dirigida) é necessário que, independentemente do contexto, a relação se mantenha. Ou seja, deve ser possível repetir o experimento ou a intervenção algumas vezes, obtendo os mesmos tipos de resultados, para poder lhe atribuir a condição de impacto. Essa característica da relação de causalidade também é apontada por Baker (2000) em sua definição de avaliação de impacto. Contudo, a autora não toca na necessidade de replicabilidade:

> A intenção da avaliação de impacto é determinar mais amplamente se o programa teve os efeitos desejados nos indivíduos, domicílios e instituições e se aqueles efeitos podem ser **atribuídos à interveção do programa**. Avaliações de impacto também podem explorar consequências não previstas, positivas ou negativas, nos beneficiários. (p. 1)[5] [grifos meus]

Para Baker a avaliação de impacto não somente se preocupa em mensurar/interpretar os resultados do programa, mas analisa em que medida eles podem ser atribuídos ao programa e **somente a ele**. Nesse sentido, a avaliação de impacto é entendida, tal qual em Scriven, como a mensuração do efeito de determinada intervenção (um programa educacional, por exemplo) sobre determinado alvo, a fim de saber em que medida houve alteração na situação inicial. A diferença entre Baker e Scriven é que a primeira busca diferenciar a avaliação de impacto da avaliação de resultados.

Outro aspecto que gera diferenças na teoria que trata de avaliação de impacto refere-se ao momento em que ela é realizada, pois a terminologia também aparece relacionada ao uso prévio da avaliação, com o objetivo de prever impactos possíveis de um programa antes de sua implementação:

> *Ex-ante* ou avaliação de impacto: uma avaliação que visa prever a probabilidade de alcançar os resultados esperados de um programa ou intervenção, ou a previsão de seus efeitos inesperados. Esta é realizada antes que o programa ou a intervenção sejam formalmente aprovados ou iniciados. Exemplos comuns de avaliação *ex-ante* são avaliações de impac-

[5] "Impact evaluation is intended to determine more broadly whether the program had the desired effects on individuals, households, and institutions and whether those effects are attributable to the program intervention. Impact evaluations can also explore unintended consequences, whether positive or negative, on beneficiaries." (Baker, 2000, p. 1)

> to ambiental e/ou avaliações de impacto social e estudos de viabilidade. (Independent..., 2006) [6]

Na citação, avaliação *ex-ante* e *impact assessment* são utilizadas como sinônimos, enquanto o uso mais comum do conceito de avaliação de impacto, relativa à medida dos efeitos de determinada iniciativa, usualmente a identifica como avaliação *ex-post*. Ressalta-se, aqui, o uso de *assessment* (normalmente relacionado à avaliação de habilidades ou cognição, ou seja, à avaliação de características de **pessoas**) e *evaluation* (mais comumente relacionado à avaliação de programas, produtos, fenômenos, etc.), indistintamente, também como sinônimos.

*Impact assessment* é outra expressão usada para designar a avaliação focada em resultados ligados diretamente a determinada intervenção. Bickman (2005), por exemplo, no verbete que produziu para a *Encyclopedia of Evaluation* prefere o uso do termo *assessment*:

> Avaliação de impacto é uma avaliação focada nos resultados ou impactos de um programa, política, organização ou tecnologia. Avaliações de impacto tipicamente tentam fazer inferência causal que conecta o avaliado com o resultado. [...] Avaliação de impacto também é referenciada como resultado, impacto ou avaliação somativa. (Bickman, 2005, p. 194)[7]

A Organização para a Cooperação Econômica e o Desenvolvimento (OCDE) também reforça que "impacto" é o efeito causado, direta ou indiretamente, por uma intervenção, claramente atribuindo esse tipo de avaliação ao final do processo de implementação do programa:

> O ponto de partida é a definição de impacto do Comitê de Assistência ao Desenvolvimento (CAD), que é: efeitos de longo-prazo, positivos e negativos, primários ou secundários, produzidos por uma intervenção para o desenvolvimento,

---

[6] "Ex-ante evaluation or impact assessment: an assessment which seeks to predict the likelihood of achieving the intended results of a programme or intervention or to forecast its unintended effects. This is conducted before the programme or intervention is formally adopted or started. Common examples of ex-ante evaluation are environmental and/or social impact assessments and feasibility studies". (Independent..., 2006)

[7] "Impact assessment is an evaluation focused on the outcomes or impact of a program, policy, organization, or technology. Impact assessments typically try to make a causal inference that connects the evaluand with an outcome. [...] Impact assessment is also referred to as outcome, impact, or summative evaluation". (Bickman, 2005, p. 194)

> direta ou indiretamente, intencional ou involuntariamente. Esta definição amplia avaliação de impacto para além de efeitos diretos para incluir a gama completa de impactos em todos os níveis da cadeia de resultados. (OECD, 2008)[8]

Observa-se nessa definição o uso do termo "impacto" relacionado ao momento em que se dá a avaliação, sendo comum na literatura o uso da palavra resultados (*outcomes*) associada a efeitos de curto e médio prazos, enquanto avaliação de impacto é associada aos resultados de longo termo, e, portanto, vai além de avaliar apenas o que aconteceu após uma intervenção (Cohen; Franco, 2008).

Nessa altura, o leitor pode estar se questionando se há realmente diferenciação entre avaliações de impactos e avaliações de resultados, ou se a questão é apenas semântica, pois a ambas é atribuído um efeito de uma determinada intervenção.

Parece inegável que há uma questão política no uso terminológico e que o que está em jogo vai além da semântica, pois o termo "impacto" pode implicar, a depender do contexto, uma conotação muito mais forte do que apenas a utilização do termo "resultado" ou "resultado de longo prazo".

Contudo, o Banco Mundial traz uma definição que permite estabelecer uma diferenciação clara entre avaliações de resultados e de impactos:

> Embora haja debate dentro da profissão sobre a definição precisa de avaliação de impacto, o uso do termo pela NONIE[9] provém da adoção da definição de impacto do Comitê de Assistência ao Desenvolvimento (CAD) da Organização para a Cooperação Econômica e o Desenvolvimento (OCDE), como "efeitos de longo-termo positivos ou negativos, primários ou secundários, produzidos por uma intervenção em desenvolvimento, direta ou indiretamente, intencional ou não-intencional". Adotar a definição do CAD leva a um foco de duas premissas subjacentes às avaliações de impacto: (a) atribuição: as palavras "efeitos produzidos por" [...] implicam uma abordagem para avaliação de impacto que é atribuir impactos a intervenções, em vez de apenas avaliar o que

[8] "The starting point is the Development Assistance Committee (DAC) definition of "impact", which is: 'positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended'. This definition broadens impact evaluation beyond direct effects to include the full range of impacts at all levels of the results chain". (OECD, 2008)

[9] NONIE (Network of Networks on Impact Evaluation) é uma rede composta pela Rede de Avaliação da OCDE, pelo Grupo de Avaliação das Nações Unidas, pelo Grupo de Cooperação para Avaliação e pela Organização Internacional para Cooperação em Avaliação.

> aconteceu. (b) contrafactual: [...] o conhecimento sobre os impactos produzidos por uma intervenção requer uma tentativa de aferir o que teria acontecido na ausência da intervenção e a comparação com o que tem ocorrido com a implementação da intervenção. (Leeuw; Vaessen, 2009, p. 9)[10]

Nesse sentido, haveria uma diferenciação metodológica relacionada ao uso de "avaliação de resultados" (entendida como medida do que aconteceu) e "avaliação de impactos": a atribuição de causalidade e a definição de um contrafactual. Principalmente o último elemento (contrafactual) parece ser a chave, na opinião da autora, para a diferenciação entre resultados e impactos, pois é possível avaliar "resultados" (o que aconteceu após a intervenção) sem estabelecer um grupo de comparação, mas este último parece essencial para se falar em impactos. Esse será o entendimento de impacto assumido neste trabalho.

A avaliação dos efeitos que são dependentes de uma intervenção é, por sua natureza, extremamente complexa, visto que questões relativas à inferência causal estão implícitas nesse tipo de avaliação (Sulbrandt, 1993).

Ora, nas ciências biológicas e exatas, isolar o efeito de uma variável pode ser mais simples do que nas ciências sociais. Como isolar, por exemplo, o efeito de um curso na prática de um profissional, sem a possibilidade de controlar os conhecimentos que ele tinha anteriormente?

Quando se trata de avaliar efeitos de um programa sobre o ser humano, cujas ações e reações envolvem uma complexidade de fatores, é possível eliminar outras explicações que possam justificar parcialmente o resultado da avaliação? É possível desenvolver indicadores ou instrumentos de medida de resultados que isolem os aspectos que podem ter interferido no processo e reflitam diferenças no objeto que sofreu a intervenção, antes e depois de ela ter ocorrido?

---

[10] "Although there is debate within the profession about the precise definition of impact evaluation, NONIE's use of the term proceeds from its adoption of the Development Assistance Committee (DAC) of the Organization for Economic Co-operation and Development (OECD) definition of impact, as 'the positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended'. Adopting the DAC definition of impact leads to a focus on two underlying premises for impact evaluations: (a) attribution: the words 'effects produced by' [...] imply an approach to impact evaluation that is about attributing impacts to interventions, rather than just assessing what happened. (b) counterfactual: [...] knowledge about the impacts produced by an intervention requires an attempt to gauge what would have occurred in the absence of the intervention and a comparison with what has occurred with the intervention implemented." (Leeuw; Vaessen, 2009, p. 9)

Soma-se a essas questões uma outra, anterior à própria medida de resultados: há informações disponíveis sobre a implementação do programa, para que se possa controlar outros fatores que possam intervir nos resultados alcançados?

Além disso, observa-se que objetivos e metas de avaliação, muitas vezes, não são bem definidos, ou mudam constantemente, dificultando a aferição dos resultados esperados e inesperados por falta de parâmetros bem definidos para balizar essa avaliação. Assim, as alterações constantes na agenda política que, frequentemente, imprimem modificações no desenho e na implementação dos programas, dificultam a realização de avaliações de impacto, que requerem metodologias mais complexas e com mais controle sobre as variáveis, a fim de que os resultados obtidos sejam confiáveis.

Na visão de Sulbrandt (1993), os aspectos mais importantes dos programas sociais que dificultam aferições de impacto são:

> a) Os problemas estruturais que se pretende enfrentar mediante as políticas e programas que são debilmente estruturados e não podem ser definidos de maneira rigorosa.
>
> b) As políticas e programas, desenhados e aprovados pelo governo, não perseguem um objetivo único, mas sim objetivos múltiplos, às vezes inconsistentes, e suas metas, da qual emanam não somente problemas técnicos, mas também necessidades táticas para assegurar sua aprovação, são definidas de maneira ambígua.
>
> c) As metas tendem a ser redefinidas no transcurso da implementação. Uma das razões que explicam estas modificações e mudanças de metas é o processo de aprendizagem social que uma organização experimenta ao desenvolver um programa.
>
> d) O caráter fraco das tecnologias utilizadas na quase totalidade dos programas sociais significa que as supostas relações causais, que vinculam os insumos e as atividades com os produtos, resultados e impactos, não respondem a um conhecimento certo e válido, mas sim que, no melhor dos casos, constituem somente hipóteses a verificar. (p. 325-326)

Segundo Sulbrandt (1993) e Rossi e Freeman (1989) esses dificultadores para a realização de uma avaliação de impacto ocorrem com bastante frequência, principalmente se o programa não prevê esse tipo de avaliação e, portanto, não há a preocupação direta com fatores essenciais para sua realização.

Dentre os autores e instituições que tratam da metodologia de avaliação pertinente à análise de impactos consultados para a elaboração deste artigo, observa-se que algumas características comuns são apresentadas:

- Definição das questões essenciais da avaliação relativas aos impactos como resultados esperados e levantamento de explicações alternativas para os resultados obtidos (relativas à seleção, atrito, efeitos externos, maturação, instrumentação)[11].
- Estabelecimento de um contrafactual (o que teria acontecido com a população alvo na ausência do programa).
- Seleção aleatória dos participantes do estudo, tanto para o grupo de "tratamento" (o que receberá a intervenção) quanto para o grupo de "controle" (que propiciará a observação do contrafactual), garantindo, ao mesmo tempo, equivalência em características que podem afetar o estudo (p. ex.: mesma classe social, faixa etária, nível de escolaridade, etc.)
- Comparação dos participantes do programa antes e depois de terem recebido a intervenção, a fim de verificar se houve ganhos de acordo com os resultados esperados.
- Comparação entre os resultados do grupo de controle e do grupo dos participantes do programa, para verificar se os resultados dos participantes excedem os resultados dos que não receberam a intervenção.
- Contextualização da avaliação (Leeuw; Vaessen, 2009; Cohen; Franco, 2008; OECD, 2008; Shadish; Cook; Campbell, 2002; Weiss, 1998).

A bibliografia de referência também destaca que os desenhos de pesquisa mais adequados à aferição de impactos seriam os experimentais e quase-experimentais, principalmente os que utilizam grupo de controle e o modelo pré-teste/pós-teste (Shadish; Cook; Campbell, 2002).

Contudo, na impossibilidade de utilização desses desenhos de pesquisa, Donald Campbell alerta que a habilidade do pesquisador para excluir qualquer explicação

[11] É importante definir o que pode ter influenciado nos resultados do programa, além da intervenção propriamente dita para, por meio do desenho da avaliação, tentar superar tais fatores intervenientes nos resultados. Muitos desses fatores podem ser desvelados pelo estudo das ameaças à validade (*threats to validity*), que muitos metodologistas que se dedicam ao desenho de pesquisa e avaliação sumarizam. Apresentar e discutir essas questões foge aos objetivos deste artigo. Entretanto, o leitor interessado em aprofundar seus conhecimentos nessa temática pode consultar Shadish, Cook e Campbell (2002).

alternativa para os resultados obtidos pela intervenção é essencial para a aferição de efeitos e impactos, mais do que o desenho de pesquisa (Weiss, 1998, p. 183).

Nesse sentido, a ideia comum de que efeitos e impactos só podem ser mensurados em desenhos experimentais e quase-experimentais, nos quais o pesquisador tem mais controle sobre as variáveis, e que dificilmente são aplicados em ciências sociais, pode ser questionada, e o debate desloca-se para o desafio que o estudioso tem que enfrentar nessa área: buscar aferir impactos, usando modelos não-experimentais de pesquisa.

Observa-se, então, a necessidade de desenvolver modelos alternativos de análise de impactos de uma intervenção que considerem as informações já existentes, que sejam factíveis e possam iluminar o entendimento sobre os resultados das ações realizadas, contribuindo com a gestão pública de serviços educacionais.

Isso implica a necessidade de retomar a discussão, em educação, sobre as tecnologias de análise disponíveis para que se isolem explicações alternativas sobre os resultados dos programas implementados como, por exemplos, ações de formação docente. Mas, para isso, é preciso também debater, com os formuladores de políticas, quais os cuidados necessários, já na implementação do programa, para que informações essenciais às avaliações de impacto possam ser produzidas ou coletadas.

Além disso, ante as dificuldades de mensurar impactos desse tipo de programa, cabe a discussão sobre como potencializar o uso das informações obtidas pelos sistemas de avaliação já existentes, visando a analisar e compreender a realidade educacional em sua complexidade e possibilitando a proposição de políticas baseadas em dados confiáveis.

Discutir as experiências de avaliação de impactos que têm sido geradas no âmbito das universidades e das instituições de pesquisa especializadas em avaliação pode servir para iluminar os meandros metodológicos da medida de impactos de programas de formação e a discussão sobre possibilidades e limites de estudar impactos sem utilizar métodos experimentais ou quase-experimentais. Tais preocupações motivam o compartilhar dos percalços gerados pela pesquisa até o presente momento, pois a busca de soluções para essa questão deveria, antes de tudo, ser coletiva.

## (DES)CAMINHOS DA PESQUISA: HÁ LUZ NO FIM DO TÚNEL?

No que se refere à formação docente, um estudo bibliográfico inicial mostrou que apesar da década de 1990 ser marcada pela proliferação de estudos sobre a formação dos professores e seu caráter de desenvolvimento profissional, observa-se que

o investimento em programas de formação continuada não parece estar contribuindo, como esperado pelos elaboradores de políticas e programas educacionais, para a melhoria da qualidade de ensino (Navarro, 2003).

É importante destacar que a existência da relação direta entre formação docente e desempenho dos alunos gera muitas controvérsias entre os pesquisadores e estudiosos, e nem sempre é aceita pela comunidade científica. Enquanto alguns autores acreditam que a relação entre a formação dos professores e o desempenho dos alunos é frágil (Torres, 1998), outros defendem que esses elementos estão intimamente relacionados (Brunner, 2003; Castro, s/d).

A análise de Marta Sisson de Castro (s/d), por exemplo, aponta a relação direta entre a formação dos professores, em nível superior, e os resultados dos alunos no PISA:

> Os resultados do PISA também constataram: "o conjunto de fatores escolares explica 31% da variância na leitura" (PISA, 2002). Ao identificar os fatores escolares que influenciam positivamente o rendimento acadêmico dos alunos, enfatizam que **professores qualificados são os recursos escolares mais valiosos.** Foi constatada uma associação entre a percentagem de professores que possuíam curso superior em sua área de atuação e resultado acadêmico dos alunos; por exemplo, uma elevação de 25% no percentual de professores com curso superior em sua área de atuação está associado com um aumento de nove pontos no teste de leitura, em média, nos países da Organization for Economic Co-operation and Development (OECD), indicando que a **preparação dos professores afeta diretamente o rendimento dos alunos.** [grifos meus]

Tal relação precisa ser tematizada por estudos que se dediquem à compreensão da política educacional, pois enquanto não são traçadas conclusões mais precisas ela não pode ser descartada como um dos elementos explicativos do sucesso ou fracasso de determinado programa, nem, tampouco, ser tomada como verdade absoluta. No caso específico da análise sobre a influência do Programa Letra e Vida nos desempenhos de alunos e professores, a intenção da pesquisadora é buscar informações que possam contribuir para o avanço da discussão dessa polêmica.

Heraldo Vianna ensina que é por meio da avaliação de um programa, aliado à pesquisa, que será possível desvendar a "rede de fatores confluentes e que se interpenetram, gerando uma rede de causas, fatos e efeitos" que interferem na realidade educacional e, portanto, na qualidade em educação (Vianna, 2005, p. 23).

Feitas tais ressalvas, e com base nas análises iniciais da autora, que tomou os resultados obtidos pelos alunos no Sistema de Avaliação de Rendimento Escolar de São Paulo (Saresp) em 2007 como indício do desempenho discente, não é possível

afirmar que tais resultados se alteraram em razão do Programa Letra e Vida oferecido aos professores de 1ª série, desde que foi implantado.

Vale destacar que a relação entre a prova do Saresp e os pressupostos teórico-metodológicos divulgados no curso existe e é colocada com clareza em documento retirado do *site* da Secretaria da Educação, à época da inscrição das escolas no Saresp, no qual são explicitadas as matrizes de referência que embasam a elaboração da prova de 1ª e 2ª séries:

> A avaliação das primeiras séries do Ensino Fundamental está vinculada à existência de professores nas redes municipal e particular que participaram do Programa de Formação de Alfabetizadores (PROFA)[12], ministrado pelo Ministério da Educação, ou do Projeto Letra e Vida, em desenvolvimento pela Secretaria da Educação do Estado de São Paulo. Essa decisão se justifica em razão da especificidade da avaliação das 1ª e 2ª séries na rede da SEE que, vinculada aos pressupostos desse Projeto, requer procedimentos específicos para a aplicação e correção de provas. (São Paulo, s/d)

Isso porque não é possível identificar, nos resultados do Saresp, uma continuidade metodológica e temporal que permita fazer afirmações fidedignas a esse respeito.

Implantado a partir de 1996, com periodicidade irregular, o Saresp já realizou dez avaliações nas escolas da rede estadual de São Paulo (Quadro 1), inclusive, em alguns anos, houve a participação de algumas redes municipais e escolas particulares.

**Quadro 1** – Edições do Saresp e séries avaliadas, por edição

| Séries / Ano | Ensino fundamental | | | | | | | | Ensino médio | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1ª | 2ª | 3ª | 4ª | 5ª | 6ª | 7ª | 8ª | 1ª | 2ª | 3ª |
| 1996 | | | ■ | | | | ■ | | | | |
| 1997 | | | | ■ | | | | ■ | | | |
| 1998 | | | | | ■ | | | | ■ | | |
| 2000 | | | | | ■ | | ■ | | | | ■ |
| 2001 | | | | ■ | | | | ■ | | | |
| 2002 | | | | ■ | | | | ■ | | | |
| 2003 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 2004 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 2005 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 2007 | ■ | ■ | | ■ | | ■ | | ■ | | | ■ |
| 2008 | | ■ | | ■ | | ■ | | ■ | | | ■ |

Fonte: São Paulo, 2005.

[12] O Programa Letra e Vida é o mesmo Programa de Formação de Alfabetizadores (PROFA) que havia sido implantado em 2001 pelo Ministério da Educação. Contudo, na experiência do PROFA a adesão dos municípios era voluntária. Os pressupostos teóricos e metodológicos, no entanto, são os mesmos, e observam-se poucas diferenças no material utilizado nos dois cursos.

Para o interesse específico da pesquisadora, observa-se que os alunos da 1ª série do ensino fundamental foram avaliados em 2003, 2004, 2005 e 2007. O Programa Letra e Vida foi implantado em 2003, o que permitiria supor que os resultados de 2004, 2005 e 2007 pudessem sofrer alterações em comparação aos de 2003 devido à ação de formação docente.

Contudo, não é possível realizar uma análise longitudinal dos resultados da 1ª série. Em 2003, por exemplo, ela foi qualitativa, não sendo atribuídas porcentagens médias para os acertos dos alunos, como mostra o trecho do Relatório do Saresp 2003:

> Cabe ressaltar que as provas das 1ª e 2ª séries do Ensino Fundamental foram corrigidas de forma qualitativa, sendo criadas categorias de desempenho; portanto, nessas séries não serão discutidas as porcentagens médias de acertos em cada habilidade, mas a porcentagem de alunos em cada categoria, por tópico de análise. As categorias de classificação das respostas dessas duas séries foram determinadas por especialistas da SEE/SP. (São Paulo, 2003)

Em 2003, os alunos foram distribuídos em níveis, segundo uma escala de desempenho como mostra o quadro 2.

Observa-se que a análise dos resultados manteve-se qualitativa, mas, diferentemente do ano anterior, associou-se um total de pontos a cada nível de desempenho.

Houve, assim, uma alteração no tratamento dos resultados que dificultou o estudo mais direto dos desempenhos das duas avaliações. Tais análises poderiam ser feitas sobre os resultados qualitativos, presentes nas duas amostras, mas as mudanças ocorridas nos descritores qualitativos poderiam levar a conclusões errôneas a respeito dos resultados e de sua relação com o programa de formação avaliado. Além disso, como não ocorreu um controle específico da questão da formação do professor, não foi possível distinguir, entre os avaliados, os que eram e os que não eram alunos de professores que haviam participado do programa.

Finalmente, a pesquisadora também perdeu a possibilidade de análise e utilização dos resultados da 1ª série, obtidos em 2005 e 2007, visto que houve alteração significativa no esquema de pontuação da avaliação, e também mudanças na disposição das categorias qualitativas.

**Quadro 2** – Níveis da escala de desempenho do Saresp 2003 para a 1ª série do ensino fundamental

| Níveis da escala de desempenho em leitura e escrita | Ensino fundamental – ciclo I – 1ª série |
|---|---|
| NÍVEL DE DESEMPENHO: **INSUFICIENTE** *(de 0 a 9 pontos)* | Neste nível, os alunos ainda não escrevem com correspondência sonora alfabética. |
| NÍVEL DE DESEMPENHO: **REGULAR** *(de 10 a 12 pontos)* | Neste nível, os alunos escrevem com correspondência sonora alfabética. |
| NÍVEL DE DESEMPENHO: **BOM** *(de 13 a 18 pontos)* | Neste nível, os alunos escrevem com correspondência sonora alfabética e leem com autonomia, localizando parcialmente informações no texto. |
| NÍVEL DE DESEMPENHO: **MUITO BOM** *(de 19 a 20 pontos)* | Neste nível, os alunos escrevem alfabeticamente com ortografia regular e leem com autonomia, localizando integralmente informações no texto e sendo capazes de inferir uma informação a partir da leitura. |
| NÍVEL DE DESEMPENHO: **ÓTIMO** *(de 21 a 24 pontos)* | Neste nível, os alunos escrevem alfabeticamente com ortografia regular e leem com autonomia, sendo capazes de inferir uma informação a partir da leitura. Produzem texto com algumas características de linguagem escrita e do gênero proposto (carta). |
| NÍVEL DE DESEMPENHO: **EXCELENTE** *(25 pontos)* | Neste nível, os alunos escrevem alfabeticamente com ortografia regular e leem com autonomia, sendo capazes de inferir uma informação a partir da leitura. Produzem texto com características de linguagem escrita e do gênero proposto (carta). |

Fonte: São Paulo, 2004.

Inicialmente, pensou-se que ao usar os dados do Saresp 2005 e 2007 propiciaria um estudo com pré-teste e pós-teste, a partir do qual se buscaria observar tendências de desempenho dos alunos e escolas cujos professores participaram do programa analisado. Contudo, enquanto o Saresp 2005 foi analisado de acordo com um escore máximo de 44 pontos, distribuídos em 8 níveis de escala de desempenho, no Saresp 2007 foram adotados 6 níveis de escala de desempenho, com uma pontuação máxima de 49 pontos, como mostram os quadros 3 e 4.

**Quadro 3** – Níveis da escala de desempenho em leitura e escrita na 1ª série do ensino fundamental. Saresp 2005

| Nível | Pontuação | Descrição dos Níveis |
|---|---|---|
| Abaixo do nível 1 | 0-4 | Alunos que não demonstram domínio das habilidades avaliadas pelos itens da prova. |
| 1 | 5-9 | Escrevem com correspondência sonora ainda não alfabética. |
| 2 | 10-12 | Escrevem com correspondência sonora alfabética. |
| 3 | 13-18 | Escrevem com correspondência sonora alfabética e leem com autonomia (texto informativo). |
| 4 | 19-25 | Escrevem com ortografia regular. |
| 5 | 26-38 | Produzem texto com algumas características de linguagem escrita e do gênero proposto (conto). |
| 6 | 39-40 | Produzem texto com características de linguagem escrita e do gênero proposto (conto). |
| 7 | 41-44 | Produzem texto com características de linguagem escrita e do gênero proposto (texto informativo), a partir de situação de leitura autônoma e de texto de outro gênero. |

Fonte: FDE, 2008.

**Quadro 4** – Níveis da escala de desempenho em leitura e escrita na 1ª série do ensino fundamental. Saresp 2007

| Nível | Pontuação | Descrição dos Níveis |
|---|---|---|
| 1 | 0-3 | Os alunos escrevem sem correspondência sonora. |
| 2 | 4-8 | Os alunos escrevem com correspondência sonora ainda não alfabética. |
| 3 | 9-16 | Os alunos escrevem com correspondência sonora alfabética. |
| 4 | 17-25 | Os alunos escrevem com correspondência sonora alfabética e produzem texto com algumas características da linguagem escrita e do gênero proposto (carta). |
| 5 | 26-37 | Os alunos escrevem com ortografia regular; produzem texto com características da linguagem escrita e do gênero proposto (carta); e, localizam, na leitura, informações explícitas contidas no texto informativo. |
| 6 | 38-49 | Os alunos escrevem com ortografia regular; produzem texto com características da linguagem escrita e do gênero proposto (carta); localizam informações explícitas; e fazem inferência de informações a partir de um texto lido (texto informativo). |

Fonte: FDE, 2008.

Observa-se que a própria distribuição dos alunos nos níveis, feita pela Secretaria da Educação e seus assessores, parece deficitária, pois os níveis são distintos e definidos com base no número de pontos variável. Independentemente da precisão desses níveis, observa-se que a diferenciação entre os instrumentos e as formas de análise trazem, em seu bojo, questões de instrumentação que, como explicado por Shadish, Cook e Campbell, consistem em uma ameaça à validade dos resultados da pesquisa caso o estudioso não consiga encontrar outras formas de análise, que não uma comparação direta entre os resultados das diversas avaliações.

A fim de tentar sobrepujar esses desafios impostos pela instrumentação, a pesquisadora optou por fazer uma análise exploratória, convertendo os resultados dos alunos em proporções. Em um exercício de reflexão, visando a encontrar pontos em comum entre os grupos que foram avaliados em 2005 e 2007, tentou-se considerar os alunos com pontuação igual ou maior que 75% nas duas provas e, para complemento do estudo, aqueles que atingiram menos de 25% da pontuação possível.

No caso de 2005, esses alunos eram aqueles que tiveram escores menores que 11 pontos e maiores que 33. Já no caso dos alunos avaliados em 2007, foram considerados os que obtiveram pontos acima de 36,75 e abaixo de 12,25. Partindo dessa análise inicial, não foi possível observar alterações nos desempenhos de alunos cujos professores participaram do curso de formação em alfabetização, que é objeto de estudo, e novas possibilidades de análise ainda estão sendo estudadas. Pode-se afirmar, contudo, que os resultados da análise exploratória não permitiram chegar a conclusões que evidenciassem impactos do curso sobre o desempenho discente, ao contrário do que se propagou à época.

Isso não significa, contudo, que o Programa Letra e Vida não alcançou resultados positivos, visto que ele pode ter atingido seus objetivos com relação à formação de professores. Entretanto, do ponto de vista quantitativo, a formação não pareceu repercutir em diferenças significativas no aprendizado dos alunos. Tal descoberta reforçou a necessidade de investigar a prática docente dos que fizeram o curso, a fim de analisar em que medida ela seria influenciada pelos pressupostos aprendidos durante a atividade de formação continuada.

Outro fator que influenciou o delineamento da análise proposta foi a dificuldade de obtenção de informações sobre o nível socioeconômico da escola, uma vez que esse tipo de informação nem sempre é incorporado aos questionários que acompanham as avaliações sistêmicas. Ora, procurar comparar resultados entre amostras equivalentes é um princípio necessário para que o pesquisador evite que os resultados sejam inválidos por questões de seleção das amostras. Assim, para poder afirmar que os resultados

não estariam sofrendo mudanças, em razão de questões de nível socioeconômico, mas sim sendo influenciados pela melhoria no desempenho do professor, por via da formação continuada, tornou-se um desafio à pesquisadora a atribuição de uma medida de característica socioeconômica a cada escola ou aluno avaliado.

Nesse sentido, apresentou-se um dificultador: os alunos do ciclo 1 do ensino fundamental não estão aptos a responder questões de nível socioeconômico, o que justifica que os organizadores do Saresp só comecem a colher esses dados a partir da 4ª série.

Para tentar ultrapassar essa limitação no estabelecimento do nível socioeconômico da escola, optou-se, então, por obter dados parciais sobre a população atendida pelas escolas e, então, generalizá-los para toda a instituição. Como alternativa, utilizou-se os dados fornecidos pela Fundação para o Desenvolvimento da Educação (FDE), que se baseou em uma adaptação do Critério Brasil[13] para, com base nas respostas dos alunos de 4ª série, traçar o perfil socioeconômico da escola. O pressuposto assumido pela pesquisadora foi que a população atendida no entorno escolar é a mesma, estando os alunos no 1º ou no 4º ano do ensino fundamental.

Além disso, desde o início do projeto, a pesquisadora intentava verificar se haveria diferenças substanciais entre o desempenho de alunos de professores que cursaram o Programa Letra e Vida e o desempenho de alunos cujos professores não participaram do curso que, como visto anteriormente, é condição essencial a uma avaliação de impacto.

Para estabelecer essa relação, seriam utilizadas as respostas dos professores ao questionário que acompanhou o Saresp de 2007 em que foram incluídas, a pedido da equipe do "Letra e Vida", questões que permitissem identificar os docentes que participaram do programa. A ideia inicial, segundo a supervisora do programa, professora Telma Weisz[14], era identificar os professores formados pelo "Letra e Vida" e cruzar essa informação com os resultados obtidos por seus alunos.

Contudo, durante a aplicação do Saresp, houve uma troca de professores aplicadores entre escolas, com exceção dos professores das duas séries iniciais do ensino fundamental. Nesse caso, os professores de 1ª e 2ª séries que aplicaram a avaliação foram os professores da própria escola, trocando, porém, as turmas.

---

[13] O Critério Brasil incorpora a escolaridade da mãe e a posse de bens de conforto, mas não questões específicas sobre renda.

[14] Informação obtida em entrevista concedida à pesquisadora em 2007.

Isso fez com que os questionários dos professores de 1ª e 2ª séries não fossem respondidos pelos responsáveis de cada turma, perdendo-se a possibilidade de cruzar os dados obtidos pelos respondentes que fizeram o "Letra e Vida" com os percentuais de rendimento dos alunos, que seria uma fonte importante de informação para a análise do impacto do programa. Novamente, para tentar superar esse problema, a pesquisadora optou por trabalhar com a escola como um todo enquanto unidade de pesquisa, e não mais com os professores individualmente. Na atual fase da pesquisa, busca-se encontrar informações sobre a quantidade de professores que tinham participado do programa em 2007 e sua composição em cada escola, em termos de proporção. Procurar-se-á, nesse sentido, estabelecer comparações entre escolas com grande porcentagem de professores que participaram da formação, e escolas com poucos professores que fizeram o "Letra e Vida", a fim de observar se há diferenças entre os resultados obtidos pelas instituições que pertencem à mesma faixa socioeconômica.

Finalmente, cabe destacar que a proposta inicial de análise de possíveis impactos do Programa Letra e Vida contemplava o acompanhamento de um grupo de professores que foram cursistas do programa, comparando os resultados dos seus alunos com os de alunos de professores que não se submeteram a ele. Tal abordagem, contudo, foi dificultada porque ocorreram inúmeras remoções de docentes durante o ano, o que não garante que, ao chegar à escola, a pesquisadora tenha acesso ao mesmo grupo que lá estava em 2007, ano em que se baseiam as informações obtidas por ela por meio da FDE.

As dificuldades encontradas no decorrer da pesquisa permitem, desde já, chegar a duas conclusões principais. Primeiramente, percebe-se que a própria natureza do sistema educacional de São Paulo dificulta a análise de resultados baseada em uma metodologia que exige o controle de variáveis e, portanto, a manutenção de algumas estruturas propostas inicialmente.

Programas que mudam constantemente, informações que não são "controladas" pelos aplicadores do Saresp (e que poderiam ser úteis à gestão do sistema), desafios técnicos e metodológicos para manutenção da unicidade nas propostas e equivalência de resultados constituem parte dos aspectos que devem ser considerados, se o objetivo é o desenvolvimento de análises mais aprofundadas e sustentáveis, com validade interna e externa.

Em segundo lugar, chama a atenção a postura da Secretaria da Educação que poderia imprimir mudanças no sentido de propiciar condições mais favoráveis para o desenvolvimento das pesquisas em educação e para o aprimoramento da reflexão teórica sobre as análises de resultados referentes ao sistema educacional.

Paralelamente, observa-se que as mesmas dificuldades enfrentadas pela pesquisadora devem se impor, de alguma maneira, ao pessoal técnico da Secretaria, dificultando-lhes o trabalho de análise dos resultados obtidos.

De qualquer forma, um maior cuidado na organização das avaliações e no tratamento de dados sobre professores poderia ser decisivo para ajudar no desenvolvimento de uma proposta de avaliação de impacto.

Contudo, são essas mesmas dificuldades que fazem com que seja necessária a realização de estudos como o que está sendo proposto neste projeto, a fim de contribuir para o acúmulo de conhecimentos na área de avaliação de impacto de programas sociais que deve ser focalizada nos próximos anos, tendo em vista as características que a gestão de programas sociais e educacionais vêm assumindo ultimamente.

## REFERÊNCIAS BIBLIOGRÁFICAS


BAKER, J. *Evaluating the impact of development projects on poverty*: a handbook for practitioners. Washington: World Bank, 2000. (Direction in development).

BAUER, A. *Usos dos resultados do Saresp*: o papel da avaliação nas políticas de formação docente. 2006. Dissertação (Mestrado em Educação) - Faculdade de Educação, USP, São Paulo.

BICKMAN, L. Impact assessment. In: MADISON, S. *Encyclopedia of evaluation*. California: SAGE, 2005, p. 194.

BRUNNER, J. J. Límites de la lectura periodística de resultados educacionales. In: UNESCO. *Evaluar las evaluaciones*: una mirada política acerca de las evaluaciones de la calidad educativa. Buenos Aires: UNESCO/IIPE, 2003, p. 67-84.

CASTRO, M. L. S. *Avaliação do rendimento educacional e a formação de professores*. Porto Alegre, [2008]. Disponível em: <http://www.sbec.org.br/evt2008/trab28.pdf>. Acesso em: 28 nov. 2008.

COHEN, E.; FRANCO, R. *Avaliação de projetos sociais*. Petrópolis: Vozes, 2008.

INDEPENDENT EVALUATORS' WEBRING. *Definitions of evaluation types, approaches and fields*. Disponível em: <http://www.evaluators-webring.net/Independent_evaluators_webring_definitions_May06.pdf>. Acesso em: 20 nov. 2008. Version as at May 2006.

LAVILLE, C.; DIONNE, J. *A Construção do saber:* manual de metodologia de pesquisa em ciências humanas. Porto Alegre: Artmed; Belo Horizonte: UFMG, 1999.

LEEUW, F.; VAESSEN, J. *Impact evaluations and development*: NONIE'S guidance on impact evaluation. Washington: World Bank, 2009.

MADISON, S. *Encyclopedia of evaluation*. California: SAGE, 2005.

MOHR, L. *Impact analysis for program evaluation*. California: SAGE, 1992.

NAVARRO, J. C. La Evaluación y las actitudes de los docentes frente a ella: dificultades y alternativas de política. In: UNESCO. *Evaluar las evaluaciones*: una mirada política acerca de las evaluaciones de la calidad educativa. Buenos Aires: Unesco/IIPE, 2003. p. 147-164.

NATIONAL SCIENCE FOUNDATION. *An Overview of quantitative and qualitative data collection methods*. Disponível em: <http://www.nsf.gov>. Acesso em: 15 nov. 2007.

OECD. *Draft NONIE statement on impact evaluation*. In: MEETING OF THE DAC NETWORK ON DEVELOPMENT EVALUATION, 7., 20-21 Feb. 2008. [S.l.]. Disponível em: <http://www.oecd.org/dataoecd/19/29/40104352.pdf>. Acesso em: 17 nov. 2008.

_____. *Outline of principles of impact evaluation*. Disponível em: <http://www.oecd.org/dataoecd/46/16/37671602.pdf.> Acesso em: 19 mar. 2010.

ROSSI, P.; FREEMAN, H. Monitoreo del programa para su evaluación. *Evaluación*: un enfoque sistemático para programas sociales. México: Trillas, 1989.

SÃO PAULO (Estado). Secretaria da Educação. *Condições de adesão da rede municipal e particular*. São Paulo, [S.d.]. Disponível em: <http://saresp.edunet.sp.gov.br/2004/subpages/condi_ad_mu.htm>. Acesso em: 21 abr. 2009.

_____. *Conhecendo os resultados do Saresp 2003*. São Paulo: FDE, 2005.

_____. *Níveis da escala de desempenho em leitura e escrita*: ensino fundamental – ciclo I – 1ª e 2ª séries. 2004. Disponível em: <http://www.educacao.sp.gov.br/noticias_2005/01_02_EF.pdf>. Acesso em: 21 abr. 2009.

_____ . *Política educacional da Secretaria da Educação do Estado de São Paulo*, 2003. Disponível em: <http://www.crmariocovas.sp.gov.br/pdf/ors/PoliticaSEE.pdf.> Acesso em: 20 abr. 2009.

_____ . *Sumário executivo do Saresp 2005 e 2007*. São Paulo: FDE, 2008. Disponível em: <http://www.educacao.sp.gov.br/saresp>. Acesso em: 20 abr. 2009.

SCRIVEN, M. *Evaluation thesaurus*. California: SAGE, 1991.

SHADISH, W.; COOK, T.; CAMPBELL, D. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Brooks/Cole, 2002.

STUFFLEBEAM, D.; WEBSTER, W. An analysis of alternative approaches to evaluation. *Educational Evaluation and Policy Analysis*, California, v. 2, n. 3, May/Jun. 1980.

SULBRANDT, J. *La Evaluación de los programas sociales*: una perspectiva crítica de los modelos usuales. Caracas: CLAD, 1993, p. 309-350.

TORRES, R. M. Tendências da formação docente nos anos 90. In: WARDE, M. (Org.). *Novas políticas educacionais*: críticas e perspectivas. São Paulo: PUC-SP, 1998. p. 173-191.

VIANNA, H. M. *Fundamentos de um programa de avaliação educacional*. Brasília: Líber Livro, 2005.

WEISS, C. *Evaluation*: methods for studying program and policies. 2th ed. New Jersey: Prentice Hall, 1998.

Recebido em: outubro 2009
Aprovado para publicação em: abril 2010