Noviembre 2019 - ISSN: 1988-7833

AVALIAÇÃO DE POLÍTICAS SOCIAIS: UMA REVISÃO BIBLIOGRÁFICA COM ENFOQUE NA ÁREA EDUCACIONAL.

Gustavo Joaquim Lisboa[1]

Para citar este artículo puede utilizar el siguiente formato:

Gustavo Joaquim Lisboa (2019): "Avaliação de políticas sociais: uma revisão bibliográfica com enfoque na área educacional", Revista Contribuciones a las Ciencias Sociales, (noviembre 2019). En línea:

https://www.eumed.net/rev/cccss/2019/11/avaliacao-politicas-sociais.html

RESUMO

A metodologia de avaliação é parte integrante dos ciclos de políticas públicas ao redor do mundo. Suas concepções e abordagens acabam definindo os parâmetros a serem adotados na análise das políticas de uma forma geral e seus resultados influenciam o desenvolvimento das ações e estratégias de programas sociais. O objetivo desse artigo é apresentar uma revisão bibliográfica contextualizando a avaliação das políticas públicas, com enfoque nas políticas sociais, especificamente na área educacional, tomando como base o Brasil. Trata-se de uma investigação bibliográfica, de análise qualitativa de conteúdos revisitados em diversos autores da área, considerando questões relativas a conceitos e objetivos da avaliação, atores, tipos e enfoques de programas avaliativos, além da discussão relativa aos experimentos em políticas sociais. Constatou-se que, não obstante a área de estudo ser relativamente nova em muitas partes do mundo, o tema tem sido alvo de constantes atualizações teóricas a partir da necessidade premente de se avaliar programas sociais. Verificou-se também a necessidade de uma discussão científica mais aprofundada acerca das avaliações dos sistemas públicos (compostos por atendimentos nas áreas sociais da educação, saúde, assistência sócia, etc.), uma vez que as metodologias utilizadas aplicam-se a programas e projetos sociais, não fazendo, na maioria das vezes, menção ao conjunto sistêmico da política.
Palavras-Chave: Avaliação de políticas públicas – Políticas sociais – Sistemas Educacionais – Ciclos de política pública.

EVALUACIÓN DE LAS POLÍTICAS SOCIALES: UNA REVISIÓN BIBLIOGRÁFICA ENFOCADA EN EL ÁREA EDUCATIVA.

RESUMEN

La metodología de evaluación es una parte integral de los ciclos de políticas públicas en todo el mundo. Sus concepciones y enfoques definen los parámetros que deben adoptarse en el análisis de las políticas en general y sus resultados influyen en el desarrollo de acciones y estrategias de los programas sociales. El objetivo de este artículo es presentar una revisión bibliográfica que contextualice la evaluación de las políticas públicas, centrándose en las políticas sociales, específicamente en el área educativa, con base en Brasil. Esta es una investigación bibliográfica, análisis cualitativo del contenido revisado por varios autores en el

[1] Doutor em Ciências, Políticas Públicas, Estratégias e Desenvolvimento pela Universidade Federal do Rio de Janeiro - UFRJ. Mestre em Desenvolvimento Regional e Meio Ambiente pela Universidade Estadual de Santa Cruz – UESC. Professor adjunto do Departamento de Economia da UESC. Endereço eletrônico: gustavo_lisboa@uesc.br.

campo, considerando temas relacionados con conceptos y objetivos de evaluación, actores, tipos y enfoques de programas evaluativos, y discusión de experimentos en políticas sociales. Aunque el área de estudio es relativamente nueva en muchas partes del mundo, ha sido objeto de constantes actualizaciones teóricas basadas en la urgente necesidad de evaluar los programas sociales. También era necesario un debate científico más profundo sobre las evaluaciones de los sistemas públicos (compuesto de asistencia en las áreas sociales de educación, salud, asistencia social, etc.), ya que las metodologías utilizadas se aplican a los programas y proyectos, más a menudo sin mencionar el conjunto de políticas sistémicas.
Palabras clave: Evaluación de políticas públicas - Políticas sociales - Sistemas educativos - Ciclos de políticas públicas.

EVALUATION OF SOCIAL PUBLIC POLICIES: A BIBLIOGRAPHIC REVIEW FOCUSING ON EDUCATIONAL AREA.

ABSTRACT

The evaluation methodology is an integral part of public policy cycles around the world. Their conceptions and approaches define the parameters to be adopted in the analysis of policies in general and their results influence the development of actions and strategies of social programs. The aim of this article is to present a bibliographical review contextualizing the evaluation of public policies, focusing on social policies, specifically in the educational area, based on Brazil. This is a bibliographic investigation, qualitative analysis of content revisited by several authors in the field, considering issues related to concepts and objectives of evaluation, actors, types and approaches of evaluative programs, and discussion of experiments in social policies. It was found that, although the area of study is relatively new in many parts of the world, the subject has been the subject of constant theoretical updates from the urgent need to evaluate social programs. There was also a need for a deeper scientific discussion about evaluations of public systems (composed of attendances in the social areas of education, health, social assistance, etc.), since the methodologies used apply to programs and projects, most often without mentioning the systemic set of politics.
Keywords: Public policy evaluation - Social policies - Educational systems - Public policy cycles.

1. INTRODUÇÃO

Avaliar tem sido o objeto de muitos teóricos ao longo dos últimos anos, sobretudo a partir dos anos subsequentes à segunda guerra mundial. Programas de avaliação têm sido criados em toda parte para tentar obter respostas não somente relacionadas aos governos, mas também à iniciativa privada.

Isso não significa que não ocorriam tentativas de avaliação de programas sociais antes desse período. No final dos anos de 1950, os programas de avaliação tornaram-se comuns em áreas como delinquência juvenil, tratamentos psicoterapeutas e psicofarmacológicos, habitação popular, atividades educacionais, iniciativas de organização comunitária, dentre outros. Durante os anos 1960, foram publicados inúmeros livros e artigos sobre o tema (ROSSI, LIPSEY e FREEMAN, 2003, p.9). Cano (2006, p. 10) corrobora afirmando que “a avaliação de programas sociais cresceu de forma notável [...] a partir dos anos 60. Surgiram teóricos da avaliação, metodólogos sobre avaliação, programas universitários dedicados a ela e associações profissionais de avaliadores.”

Durante os anos de 1970 e 1980 os programas de pesquisa em avaliação passaram a emergir como uma especialização no campo das ciências sociais. Com o desenvolvimento das pesquisas e a elaboração de métodos sofisticados de análise, os governos e as empresas passaram a financiar programas de avaliação. Hodiernamente, existe vasto campo teórico sobre o tema e a avaliação tornou-se parte integral da política social e dos movimentos da administração pública. (ROSSI, LIPSEY e FREEMAN, 2003, p. 11)

A tarefa de avaliar, contudo, requer esforço metodológico que viabilize os resultados de um dado programa, fornecendo informações que possuam credibilidade e segurança. Esse

processo é um dos ciclos da política pública[2], que possui grande importância para a determinação do alcance de resultados, metas e objetivos de determinados intentos.

Este estudo será dedicado à apresentação da justificação, conceitos, objetivos, tipos e enfoques da avaliação de políticas públicas, com forte apelo aos programas sociais, sobretudo na área de educação. Serão discutidas tanto avaliações de programas quanto sistêmicas, que embora possuam a mesma raiz metodológica, diferenciam-se nos desenhos implementados.

2. JUSTIFICAÇÃO, CONCEITOS E OBJETIVOS

Antes, no decurso, ou mesmo depois de uma política pública ser implementada, muitas questões precisam ser esclarecidas para os *policymakers, decisionmakers*, *policytakers*[3], *stakeholders* e a comunidade em geral.

Em arenas públicas mais descentralizadas e democráticas, cujo ideário da participação se efetiva com menores atritos de interesse, audiências públicas, podem representar um ganho na identificação de atores que, por sua vez, acabam por contribuir na formulação de políticas que possuam maior viabilidade e atendam aos interesses do público alvo com efetividade.

Na prática, *policymakers* e *decisionmakers* estariam preocupados em saber se o programa atingirá seus objetivos e metas com os menores custos possíveis, os *policytakers* estariam interessados em obter o maior benefício quanto possível do programa e a comunidade como um todo, de certa forma, estaria interessada nos resultados e no uso dos recursos públicos na implementação do mesmo.

As respostas decorrentes de tais questionamentos nem sempre são simples, até porque muitas delas não possuem clareza *ex-ante*, muito embora existam técnicas que ajudam a traçar um panorama capaz de diagnosticar os resultados de um dado programa. Além disso, mesmo as práticas utilizadas para avaliar resultados de programas já concluídos amoldam-se a técnicas de avaliações que visam reduzir os efeitos de outros fatores sobre os *policytakers*, no sentido de estabelecer relações confiáveis entre diversas variáveis de análise.

Por esse motivo, implementar políticas impõe saber qual o nível da necessidade do público alvo antes de fazer parte de uma dada política e como mensurar as mudanças ocorridas a partir das ações e atividades programadas no decurso da mesma. Rossi, Lipsey e Freeman (2003, p. 3) justificam programas de avaliação[4] a partir da definição de algumas questões básicas:

> *What are the nature and scope of the problem? […] What is it about the problem or its effects that justifies new, expanded, or modified social programs? What feasible interventions are likely to significantly ameliorate the problem? What are the appropriate target populations for intervention? […] Is the intervention being implemented well? […] Is the intervention effective in attaining the desired goals or benefits? Is the program cost reasonable in relation to its effectiveness and benefits?*

Uma questão crucial apresentada pelos autores é a necessária avaliação para que se tenha clareza acerca dos resultados dos intentos de uma política, visando à continuidade ou não das ações contempladas, caso o programa seja continuado ao longo do tempo, ou mesmo sua extinção. O que se observa nos casos em que os resultados não são mensurados adequadamente é a proliferação de programas muitas vezes ineficazes e dispendiosos que, sem a devida verificação, cumprem somente função da agenda pública de um governo, sem apresentar melhorias e avanços relativos aos *policytakers*.

Os programas de avaliação concebidos dessa forma contribuem, por meio de métodos de pesquisa social, com a tarefa de melhorar os desenhos, a implementação, o impacto e a eficiência de ações públicas que focam problemas específicos e, na medida em que se avolumam as pesquisas dos programas, acumulam-se conhecimentos sobre cada estudo individual que podem ser contribuições vitais para consubstanciar ações sociais, ajudando a melhorar as condições humanas (ROSSI et al., 2003).

Além disso, é importante verificar a natureza das mudanças, conforme alerta Ridge (2010, p. 11) *"[...] Programs resulted in conditions being diferente, but were they better? If they*

---

[2] De acordo com Secchi, (2013, p. 43), o ciclo de políticas públicas (*policy cycle*) pode ser definido como "um esquema de visualização e interpretação que organiza a vida de uma política pública em fases sequenciais e interdependentes". O autor afirma que o ciclo é composto das seguintes fases: identificação do problema, formação da agenda, formulação de alternativas, tomada de decisão, implementação, avaliação e extinção.

[3] São os usuários diretos do serviço público, o alvo para os qual é direcionada uma política pública.

[4] Para esses autores, um programa de avaliação, ou pesquisa avaliativa, é composto de coleta, análise, interpretação e informação sobre a efetividade dos programas sociais.

were better, was the improvement worth the investiment? Could the same change take place with fewer resources?"

Em todas as questões discutidas, a necessidade de concepção de um programa social é latente a partir da identificação de problemas públicos. Percebido assim, "[...] um programa social é uma intervenção sistemática planejada com o objetivo de atingir uma mudança na realidade social" (CANO, 2006, p. 9).

A complexidade e variedade de informações que devem ser levadas em conta pelos *decisionmakers* justificam, em grande medida, as avaliações das políticas públicas, não somente em relação aos métodos estatísticos de mensuração e comprovação de resultados, mas também em relação ao tipo de desenho de pesquisa ou tipos alternativos de intervenção.

O problema da repetência escolar, sobretudo na educação básica brasileira, por exemplo, pode ser objeto de várias políticas públicas. A formulação desses programas, entretanto, pode resultar de desenhos diversos. Pode-se promover a progressão continuada do estudante para que o mesmo atinja os níveis de aprendizagem adequados ao seu ano de escolaridade, por meio de atividades escolares e complementares no contra turno do ano seguinte, ou pode-se ampliar o atendimento a estudantes com baixo desempenho escolar a partir da identificação e seleção de estudantes no início de cada ano letivo. A escolha desses desenhos alternativos deve levar em consideração muitos fatores, tais como evasão escolar, motivação do aluno, custos e financiamento da educação básica, dentre outros.

O discorrido até aqui no sentido da justificação das avaliações, dada a natureza holística e multidisciplinar das políticas públicas, permite a introdução do conceito de avaliação de políticas públicas.

Embora não exista um único conceito de avaliação, epistemologicamente, o cerne conceitual gira sempre em torno de um mesmo objeto. Schneider (2010, p. 318) compreende que a "avaliação envolve a análise de programas ou políticas, em termos de seu nível de desempenho". Para o autor, tal desempenho pode gerar externalidades negativas ou positivas, devendo ser "determinado em termos de conceitos indicados nominalmente na legislação pertinente ou nas diretrizes, ou em relação às expectativas da clientela, ou pela identificação de prováveis consequências".

Em meados dos anos 2000, o Governo Federal Brasileiro, por meio do Ministério do Planejamento, Orçamento e Gestão tentou criar um sistema federal de avaliação a partir da construção de indicadores com base no Plano Plurianual – PPA[5] de diversas instituições federais. Teoricamente, o trabalho estava apoiado em legislação pertinente. No entanto, apoiar-se apenas em indicadores que levam em conta a quantidade de recursos utilizados ou percentuais de realização de leis pode comprometer sobremaneira a avaliação dos resultados, por não levar em conta uma série de varáveis que poderiam afetar o público alvo dos programas. Além disso, Santos (2012, p. 51-61) chegou à conclusão que tais avaliações possuíam "insuficiência na geração de informações sobre resultados; falta de capacitação dos envolvidos nos processos de monitoramento e de avaliação; e inadequação dos indicadores para a aferição dos resultados".

Rossi, Lipsey e Freeman (2003, p. 16) conceituam a avaliação como *"[...] the use of social research methods to systematically investigate the effectiveness of social intervention programs in ways that are adapted to their political and organizational enviroments and are designed to inform social action to improve social conditions".* Isso implica em julgar programas realizados por instituições diversas baseado em padrões, critérios, indicadores e índices que permitam a mensuração do desempenho de programas sociais em relação aos seus objetivos e, sobretudo, verificar as condições do problema social após as intervenções.

Diversos autores corroboram com essa linha de pensamento, a exemplo de Anderson (1979, p. 711) que considera avaliação como *"the process of making deliberate judgments on the worth of proposals for public action as well as on the sucess or failure of projects that have been put into effect"*. Do mesmo modo está o raciocínio de que a avaliação procura provar que mudanças sofridas pelos *policytakers* são exclusivamente devidas à implementação de uma específica política pública (KHANDKER, KOOLWAL e SAMAD, 2010, p. 8). De maneira ainda mais geral está o entendimento de Gertler et al.(2010, p. 7):

> *Evaluations are periodic, objective assessments of a planned, ongoing, or completed project, program, or policy. Evaluations are used to answer specific questions related to design, implementation, and results. […] Their design, method, and cost vary substantially depending on the type of question the evaluation is trying to answer.*

[5] Instrumento de planejamento do governo para os próximos quatro anos. Deve ser elaborado no primeiro ano de gestão de cada governo e é válido até o primeiro ano da próxima gestão administrativa.

Há que se fazer, ainda, algumas distinções entre avaliação e outras formas de busca de informações sobre a execução e os respectivos resultados das políticas públicas, pois a avaliação não é a única maneira de obter respostas sobre as ações públicas. Termos como avaliação, monitoramento ou acompanhamento, monitoramento de desempenho *(performance measurement)*, análise e até mesmo auditoria são comuns na arena política, e possuem diferenças significativas entre si. A UNICEF (1990, p.2) define monitoramento como o acompanhamento periódico de atividades e ações públicas, no sentido de estabelecer a medida em que os serviços fornecidos, os horários de trabalho, dentre outras questões, estão de acordo com o planejamento da política, de modo que medidas oportunas possam ser tomadas para corrigir deficiências detectadas.

A avaliação, por sua vez, é um processo que tenta determinar sistematicamente a possível relevância, eficácia, eficiência e impacto das atividades de um programa à luz dos seus objetivos. É, portanto, uma ferramenta de aprendizagem e de gestão orientada para a ação e processo organizacional, com o intuito de melhorar as atividades atuais e futuros planejamentos, programação e tomada de decisão (UNICEF, 1990, p. 2).

O monitoramento tem uma perspectiva gerencial na medida em que fornece informações sobre o *status* de um programa, ou seja, o *"[...] monitoring provides feedback to warn a social service administrator when the implementation of a program starts to deviate from its original design"*. A avaliação está preocupada em fornecer resultados e impactos do programa para políticos, tomadores de decisão e planejadores, não se preocupando com a comunicação de dados de resultados e informações a stakeholders externos (monitoramento de desempenho) ou com o acompanhamento de sucesso de um programa na realização de resultados e impactos durante sua implementação (monitoramento) (KETTNER, MORONEY e MARTIN, 2012, p. 256).

O monitoramento acompanha entradas, atividades e saídas, embora ocasionalmente possa incluir os resultados, tais como o progresso em direção a metas nacionais de desenvolvimento. As avaliações são utilizadas para responder específicas questões relacionadas com a concepção, implementação e resultados. Avaliações contrastam com os monitoramentos, pois são realizadas em pontos discretos no tempo e muitas vezes procuram uma perspectiva externa de especialistas técnicos. (GERTLER et al., 2010, p. 7)

Não obstante os conceitos serem diferentes, é perceptível a inter-relação entre os dois, vez que:

> [...] os processos de monitoramento e avaliação são complementares, porém a avaliação vai além, na medida em que realiza a verificação de que o plano originalmente traçado está efetivamente alcançando as transformações que pretendia, subsidiando a definição de políticas públicas. Mas a avaliação necessita das informações provenientes do monitoramento para realizar o julgamento que lhe cabe, a respeito da eficiência, eficácia e efetividade dos programas. (CUNHA, 2006, p. 12)

Mas qual seria a diferença entre monitoramento e monitoramento de desempenho? O *Government Accountability Office (GAO)*, dos Estados Unidos, apud Cunha, (2006, p. 12) define monitoramento de desempenho como "[...] relato contínuo do acompanhamento de um programa, particularmente de seu progresso em direção aos objetivos previamente fixados, conduzida pelo gerente do programa ou da agência responsável pelo programa". Nesse sentido, medidas de desempenho podem ser criadas sobre as diversas fases da política, inclusive seus resultados, servindo "[...] como um sistema de alerta para o gerente, em função de sua natureza contínua". Além disso, como bem destacaram Kettner, Moroney e Martin (2012, p. 256), uma das importantes funções do monitoramento de desempenho é informar *stakeholders* externos por meio de indicadores de fácil comunicação e interpretação.

Um outro tipo de análise que pode ser realizada acerca da coisa pública é a auditoria. Esta por sua vez, serve a vários usos e possui grande ramificação de conceitos e formas de atuação. Originalmente, a auditoria[6] possuía caráter fiscalizador sobre o uso dos recursos públicos, materiais e intangíveis, transformando-se em um instrumento de avaliação capaz de abranger vários aspectos de uma gestão pública, tais como recursos humanos e patrimoniais, contas públicas, compras e almoxarifado. O objetivo investigativo de sua avaliação auxiliava as decisões políticas, na medida em que além de servir como base de orientação também era capaz de antecipar cenários.

[6] Ver Félix (2008, p. 6).

Com o passar do tempo, novas técnicas foram introduzidas e seus conceitos adaptados. Hodiernamente, a auditoria passou a incluir indicadores de desempenho qualitativo (auditoria de desempenho) na avaliação de setores públicos. Diante das inúmeras possibilidades de uma auditoria, Félix (2008, p. 6-7) compreende a auditoria de desempenho dessa forma:

> Nos dias atuais os modelos de auditoria adotados podem ser denominados de auditoria contábil, de gestão, de recursos externos, de sistemas, de programas, operacional, de avaliação de metas, de conformidade, de acordo com o tipo de análise a ser efetuado. Isto porque o conceito a auditoria é mais dinâmico, atribuindo-lhe funções que abrangem toda a administração da empresa, passando a ter um regime mais voltado à orientação, interpretação e previsão de fatos [...] A Auditoria de Desempenho tem a capacidade de influenciar a administração na medida em que possibilita uma visão gerencial do sistema através da medição dos programas governamentais e da atuação dos gestores públicos, do fornecimento de sugestões para melhor alocação de recursos físicos e financeiros, na detecção de erros e fraudes, na avaliação dos controles internos, na avaliação da execução orçamentária.

No entanto, na visão de Barzelay (2014, p. 3) não existe consenso sobre o conceito de auditoria de desempenho, sendo o termo normalmente usado para diferenciá-la de auditoria tradicional e de avaliação de programas, assim:

> [...] A forma como essa distinção é feita tem implicações sobre qual categoria profissional possui os argumentos mais persuasivos com o intuito de obter o controle jurisdicional da matéria. Os funcionários de alguns órgãos centrais de auditoria tendem a destacar as semelhanças entre a auditoria tradicional e a de desempenho (Sedgwick, 1993), provavelmente com o intuito de manter ou expandir sua fatia de mercado na indústria da revisão governamental. Os especialistas em avaliação, por seu turno, tendem a caracterizar a auditoria de desempenho como uma forma de avaliação (Chelimsky, 1985; Rist, 1989), talvez pela mesma razão.

De qualquer forma, independentemente das questões meramente conceituais, a auditoria conduz a outra condição também fundamental, a da *accountability*. Com a diversidade de conselhos municipais e a difusão das tecnologias da informação, as auditorias de desempenho passam a auxiliar *policymakers* e *stakeholders* dos mais diferentes setores públicos, na identificação de problemas e fraudes na administração pública[7].

A discussão final entre avaliação de políticas públicas e outros mecanismos, dar-se-á pela distinção entre análise e avaliação de políticas públicas. Como já discutido anteriormente neste trabalho, a avaliação presta-se ao objetivo do julgamento, de estabelecer conexões entre políticas e resultados. Avaliação, vista dessa forma, "prescinde do exame da operacionalidade concreta ou da implementação do programa sob análise. Ela examina os pressupostos e fundamentos políticos de um determinado curso de ação pública, independentemente de sua engenharia institucional e de seus resultados prováveis". Por análise da política pública, "entende-se o exame da engenharia institucional e dos traços constitutivos dos programas. Qualquer política pública pode ser formulada e implementada de diversos modos" (ARRETCHE, 1998, p. 2).

Compreendida assim, analisar uma política em educação, por exemplo, pode ser contextualizá-la sob a forma de sua institucionalização, dos seus desenhos e traços, do modelo mental adotado; verificar questões ligadas à participação pública, nível de relação com iniciativa privada, adoção de determinada organização do ensino, dentre outras. Acredita-se que:

> [...] A análise de políticas públicas busca reconstituir estas diversas características, de forma a apreendê-las em um todo coerente e compreensível. [...] Ainda que a análise de uma dada política pública possa atribuir a um determinado desenho institucional alguns resultados prováveis, somente a avaliação desta política poderá atribuir uma relação de causalidade entre um programa x e um resultado y (ARRETCHE, 1998, p. 2).

Na tentativa de ampliar o debate para além das diferenças conceituais, Boschetti (2006, p. 3) defende que há inter-relação e complementariedade entre os diversos sentidos, momentos e movimentos avaliativos, dando ênfase à importância das políticas sociais como forma de consolidação do Estado democrático de direito, o que centraliza o foco das políticas

[7] Axelsen et al. (2011) desenvolveram discussões acerca da influência dos sistemas de informação *(information systems)* na auditoria financeira.

no “âmbito da identificação da concepção de Estado e de política social que determina seu resultado”.

O último conceito relativo a avaliação é a metaavaliação ou estudos de avaliação sobre avaliações. É crivo e notório que diante de extensas possibilidades alternativas, as decisões sobre os desenhos e métodos a serem aplicados em um programa de avaliação são passíveis de críticas e, igualmente, analisados sob abordagens diversas. Além disso, não somente os métodos quantitativos ou qualitativos para o alcance dos resultados estão na berlinda, mas também a condução da avaliação, se participativa, se centralizada, dentre outros critérios, levando os avaliadores a serem responsivos pelos resultados gerados a partir do conjunto da avaliação.

Henry (2001) apud Trevisan e Bellen (2008, p. 547) acredita que:

> “[...] o interesse em metaavaliações pode produzir duas linhas importantes de trabalhos para o campo. Em primeiro lugar, um resultado muito positivo poderia ser um conjunto de estudos empíricos que examinam os impactos verdadeiros das avaliações. A segunda contribuição das metaavaliações diz respeito às revisões de avaliações individuais.”

Tal inclinação pode identificar fatores de risco em avaliações que, à primeira vista, apresentam-se como soluções confiáveis à tomada de decisão e, por outro lado, contribui para a identificação de possíveis vieses propositais ocorridos em razão do financiamento da pesquisa ou por interesse das instituições e órgãos públicos que, porventura, contratem pesquisas direcionadas[8].

Uma amplitude ainda maior dada a essa questão está direcionada à política da avaliação de políticas públicas, no sentido da compreensão das formas de uso das avaliações. Estudo realizado por Faria (2005, p. 106) indica que a “preocupação com a questão do uso da pesquisa avaliativa parece ainda fortemente restrita à utilização gerencial da avaliação e à necessidade de se gerar *feedbacks* que justifiquem a relevância da própria realização de tais estudos”. O autor complementa, dizendo que:

> Uma postura de omissão no tratamento analítico das questões associadas ao processo de avaliação das políticas públicas e de seu uso, como aquela detectada no caso da ciência política brasileira (a qual, diga-se de passagem, parece não se singularizar neste aspecto), significa o esvaziamento da possibilidade de se analisar de forma cabal a política da avaliação de políticas, a qual acaba, assim, negligenciada em muitos de seus aspectos e implicações (FARIA, 2005, p. 106).

Um ponto importante é a relação entre a avaliação e todos aqueles que fazem parte do programa, sobretudo para que o uso dos programas avaliativos possam se efetivar. Esse debate é abordado na distinção entre avaliação de políticas e cultura de avaliação, apresentada por Firme, Letichevsky e Dannemann (2009, p. 172), em um estudo de caso realizado no Brasil:

> […] *“Evaluation culture” in Brazil is presently an overall comprehension and acceptance in the realm of institutions, organizations and society in general, of evaluation importance, need, practice and utilization. In other words, when “evaluation culture” is present, there is a set of learned beliefs, values, styles and behaviors that makes evaluation welcome. Evaluation here does not threaten or condemn*. […] *“evaluation policy” is perceived as a set of guidelines that establishes rules and procedures to properly conduct planning, implementation and utilization of evaluation, in all circumstances or levels of possible implementation*.

Partindo desse pressuposto, pode-se afirmar que avaliações serão tão mais participativas quanto maior for a cultura de avaliação dos *policymakers*, *decisionmakers* e *stakeholders*. Reforçando essa assertiva, existiriam três categorizações referentes à relação entre a avaliação, implementação e cultura avaliativa: *“[…] evaluation without a guiding policy; evaluation with a pre-determined policy designed by a small group, not submitted for wide and substantial discussion […] evaluation with a pre-determined policy designed in an open and collective manner, and widely disseminated”* (FIRME, LETICHEVSKY E DANNEMANN, 2009, p. 172).

No entanto, desenhos participativos levam a um esforço maior no sentido do controle de todos os interesses envolvidos no processo político, em que o avaliador deverá ser capaz de equacionar pontos de vista diferenciados entre, por exemplo, *stakeholders* e

[8] A esse respeito, ver Derlien (2001); Henry (2000); e Faria (2005).

decisionmakers. Nesses casos, pode haver forte disputa entre grupos políticos diferentes, percepções ideológicas antagônicas, ou ainda, encontrar o melhor desenho, mas percebê-lo inviável, seja por uma questão temporal, financeira ou, até mesmo, ética.

3 PROGRAMAS DE AVALIAÇÃO, CIÊNCIA E CASUALIDADE

Há que se perguntar se avaliação trata-se de ciência ou uma postura pragmática. Neste caso, Donald Campbell e Lee Cronbach, teóricos da avaliação, possuem visões diferentes sobre o tema. Campbell acredita que: *"[…] the technology of social research made it feasible to extend the experimental model to evaluation research to create an 'experimenting society' […] it is fair to characterize him as fitting evaluation research into to scientific research paradigm"* (ROSSI, LIPSEY e FREEMAN, 2003, p. 23). Não obstante, Cronbach afirma que avaliação é mais arte do que pesquisa, na medida em que:

> "*[...] the purpose of evaluation sharply differentiates it from scientific research. [...] should be orientated toward meeting the needs of program decisionmakers and stakeholders [...] should be dedicated to providing the maximally useful information that the political circumstances, program constraints, and available resources allow (ROSSI, LIPSEY e FREEMAN, 2003, p. 23).*

3.1 Atores, tipos e enfoques de programas avaliativos

Existem diversas formas diferentes de tentar encontrar resultados alcançados pelas políticas públicas. Tais avaliações podem ser aplicadas mesmo antes da implementação da política (avaliação *ex ante*) ou após a conclusão da mesma (avaliação *ex post*), ou mesmo durante o processo de execução e implementação (avaliação *in intinere*).

Pode-se estar avaliando programas novos ou já estabelecidos e concebidos há muito tempo, com objetivos distintos, por exemplo, saber se um programa é economicamente viável, ou se os benefícios sociais foram suficientes para que o poder público garanta orçamento, ou até o amplie, visando à sua manutenção.

Não se deve, no entanto, menosprezar o efeito que o contexto administrativo e político exerce sobre os programas avaliativos. Quase sempre o avaliador estará diante de situações em que não poderá determinar metas, objetivos, foco e até mesmo a metodologia da avaliação, pois estará diante de incertezas administrativas relativas a custos, financiamento, acessibilidade de informações, dentre outros.

Além disso, em uma avaliação participativa os desejos dos *policymakers*, *decisionmakers* e *stakeholders* poderão ser antagônicos, e as dificuldades impostas pela relação entre as pessoas e suas ideologias, crenças e teorias podem dificultar o processo avaliativo de modo a inviabilizá-lo. Por vezes, as valorações dos *stakeholders* estão além dos objetivos de um programa e caberá, aos responsáveis pela formulação do mesmo, atentar para esse fato, equalizando e balanceando as diversas correntes de pensamento e propondo soluções factíveis.

Tais soluções compõem a teoria do programa ou programa teórico (*program theory*). A teoria do programa é o plano de operação, a lógica que conecta as atividades com os resultados pretendidos. Por esse motivo, quanto mais complexas, descentralizadas e dispersas forem as estruturas organizacionais, maiores serão as dificuldades práticas da avaliação, o que precisa ser equacionado no plano de operação (ROSSI, LIPSEY e FREEMAN, 2003, p. 44).

Programas mais fáceis de avaliar seriam aqueles com atividades mais concretas, a exemplo de fornecimento de comida a pessoas desabrigadas, cujos resultados são especialmente imediatos e observáveis. Enquanto intervenções que são difusas na natureza (organização comunitária), se estendem por longos períodos de tempo (currículo escolar) ou precisam ser comprovadas no longo prazo (políticas compensatórias na pré-escola) possuem maior grau de complexidade e exigem maior refinamento técnico em sua formulação. Não obstante, o tempo e os prazos de uma avaliação não são determinados pelo avaliador, mas pelas condições impostas pelo contexto político sob o qual será estabelecida a análise, e quanto mais ampla, profunda e rigorosa for a avaliação maiores deverão ser seu financiamento e tempo, devendo os métodos e procedimentos, a serem adotados, levar em consideração a relação entre a técnica e o tempo (ROSSI, LIPSEY e FREEMAN, 2003, p. 46).

3.1.1 Atores do processo de avaliação

A relação interpessoal entre os atores presentes em uma avaliação e destes com o(s) avaliador(es) de uma política pública é essencial e determinante para o êxito de um programa teórico e, consequentemente, para a exitosa mensuração de resultados.

É importante, para o sucesso de um programa avaliativo, a identificação de todos os atores no processo da política pública. Tais atores são definidos por Secchi (2013, p. 99) como "aqueles indivíduos, grupos ou organizações que desempenham um papel na arena política. Os atores relevantes em um processo de política pública são aqueles que têm a capacidade de influenciar, direta ou indiretamente, o conteúdo e os resultados da política pública".

Dessa definição, ampla por assim dizer, infere-se que indivíduos estariam representando governos, famílias, organizações sociais, conselhos comunitários e de acompanhamento de programas públicos, instituições jurídicas, dentre outros, e os mesmos "não têm comportamentos ou interesses estáticos, mas sim dinâmicos de acordo com os papeis que representam" (SECCHI, 2013, p. 99). Depreende-se daí que os atores também podem ser categorizados em individuais e coletivos[9].

Corroborando com a identificação dos principais *stakeholders*, Rossi, Lipsey e Freeman (2003, p. 48) apresentam os seguintes atores: *policymakers and decisionmakers* (responsáveis pela decisão acerca do programa); *Program sponsors* (Organizações financiadoras e promotoras do programa); *Evaluation sponsors* (organizações que financiam as avaliações); *Target participants* (receptores da intervenção); *Program managers* (pessoal responsável pela administração do programa); *Program staff* (pessoal responsável pela prestação do serviço ou apoio); *Program competitors* (organizações, grupos ou indivíduos que competem com o programa no sentido da viabilização de recursos); *Contextual stakeholders* (todos que estão próximos ao ambiente do programa e desejam saber o que ele faz ou o que está acontecendo); e *evaluation and research community* (profissionais de avaliação e pesquisadores que trabalham em áreas relacionadas ao programa).

Outra categoria de atores foi definida por Moon e Ingraham (1998), em sentido mais geral ao estudarem reformas administrativas. O esquema analítico, conhecido como *Political Nexus Triad*, identificou "[...] três categorias principais: políticos (eleitos e seus designados politicamente); burocratas (selecionados por concurso) e sociedade civil (externos à administração pública)" (SECCHI, 2013, p 100).

Aplicando esse modelo à estrutura administrativa educacional de grande parte dos municípios brasileiros, por exemplo, poder-se-ia dizer que os políticos seriam representados pelos Secretários da Educação ou Gestores da Educação, e todo o *staff* das secretarias, tais como diretores, assessores e gerentes. Os burocratas seriam todos os concursados em funções de apoio e execução, tais como os profissionais do magistério, pessoal de apoio das escolas, responsáveis pela matrícula e alimentação escolar, dentre outros. A sociedade civil embora possa ser representada por diversas organizações e movimentos sociais locais, possui representação plena e regulamentada nacionalmente nos mais diversos setores educacionais, a exemplo dos conselhos de educação (questões normativas e deliberativas), dos conselhos do FUNDEB (análise e pareceres sobre contas públicas) e conselhos de alimentação escolar (acompanhamento e controle dos gastos em alimentação e qualidade da merenda escolar servida nas unidades escolares).

Modelo alternativo de análise divide atores em governamentais (políticos, designados politicamente, burocratas e juízes) e não governamentais (grupos de interesse, partidos políticos, meio de comunicação, *think tanks*[10], destinatários das políticas públicas, organizações do terceiro setor e outros *stakeholders* – fornecedores, organismos internacionais, comunidades epistêmicas, financiadores, especialistas, etc.) (SECCHI, 2013, p 101).

Nos últimos anos os atores têm assumido papel importante na formulação de programas teóricos, não somente na identificação de fatores relativos ao desenho das pesquisas, como também na participação efetiva na busca de resultados[11].

3.1.2 Critérios e tipos das avaliações de políticas públicas

[9] "[...] atores individuais são pessoas que agem intencionalmente em uma arena política [...] políticos, burocratas, os magistrados, os formadores de opinião. Atores coletivos são os grupos e as organizações que agem intencionalmente em uma arena política. [...] partidos políticos, a burocracia, os grupos de interesse, as organizações da sociedade civil e os movimentos sociais." (SECCHI, 2013, p. 100)

[10] Organizações de pesquisa e aconselhamento em politicas públicas.

[11] Ver Franco, Brooke e Alves (2008).

Como a avaliação é, de forma geral, a mensuração de resultados alcançados por uma determinada política pública, faz-se necessário decidir quais serão as formas de apresentação desses resultados, ou seja, é preciso que se encontre a melhor maneira de se chegar ao resultado. Portanto, antes da realização de qualquer procedimento relacionado a uma avaliação, algumas questões serão fundamentais:

a. Qual a natureza e o objetivo da avaliação?
b. Quais serão os critérios adotados para a mensuração dos resultados? Serão utilizados mecanismos estatísticos que gerem índices, indicadores, relações entre custos de programas, ou serão utilizados conceitos e pareceres na avaliação?
c. Quem fará a avaliação? Os procedimentos serão realizados por meio do *staff* de uma organização, ou será conduzida por técnicos externos à instituição?
d. A avaliação será feita *ex ante*, *in intinere* ou *ex post* à execução, implementação do programa?
e. O programa teórico será destinado a um projeto específico ou a um sistema ou setor público?
f. Quais problemas éticos podem ser encontrados na implementação de uma avaliação?

Tais questionamentos levam à necessidade de estabelecer critérios claros que possam favorecer o discernimento acerca do verdadeiro propósito da avaliação. Os critérios são variados e complexos, muito embora sejam quase consensuais hodiernamente. A UNICEF (1990, p. 46-47) preconiza alguns deles:

a. Efetividade: o projeto ou programa atingiu progresso satisfatório em relação aos objetivos pretendidos?
b. Eficiência: os efeitos foram alcançados a um custo aceitável, em comparação com abordagens alternativas que realizariam os mesmos objetivos?
c. Relevância: os objetivos do projeto ainda são relevantes? O problema abordado ainda existe?
d. Impacto: quais são os resultados do projeto? Quais são os efeitos socioeconômicos, ambientais, técnicos sobre os indivíduos, as comunidades e as instituições?
e. Sustentabilidade: haverá continuidade do programa, projeto ou após as etapas de implementação a política será extinta?

A partir desses critérios gerais derivam-se muitos outros, que serão adotados na medida da necessidade da avaliação em razão da natureza do projeto, programa ou sistema a ser analisado.

O momento em que se decidem os critérios dependerá do *timing* da avaliação. Nesse sentido, o tempo, ou momento de uma avaliação foi por Mendes et al. (2013, p. 99) assim conceituado:

> a) *ex ante* – fase inicial ou de predecisão [...] Pretende abarcar três aspectos: pertinência do projeto em relação à realidade; [...] diagnóstico e proposições; rentabilidade econômica das diferentes ações para alcançar os objetivos propostos. b) Avaliação "durante" a execução [*in intinere***, grifo nosso**] – busca fornecer informações sobre o andamento do programa ponderando os resultados. Avalia as mudanças situacionais, para identificar até que ponto está sendo cumprido e realizado o programa estabelecido. c) Avaliação "ex post" – realiza-se ao término do programa, chamada avaliação de impacto ou avaliação pós-decisão, visa avaliar quanto e como mudou a "situação inicial", ou quanto se alcançou a "situação objetiva", segundo o referencial traçado.

O *timing* da avaliação conduz à sua natureza: formativa ou sumativa. A primeira diz respeito à formação do programa e ocorre durante sua implementação, sendo voltada para a análise e produção de informação sobre as etapas de implementação, gerando informações para aqueles que estão diretamente envolvidos no programa. A sumativa preocupa-se com análises e produção de informações sobre etapas posteriores, produzindo informações no sentido da verificação de sua efetividade, com juízo de valor sobre seus efeitos (CUNHA, 2006, p. 10).

O conjunto dessas observações gera a tipologia da avaliação. Embora existam diversas nomenclaturas para cada tipo, cinco metodologias de avaliação são mais comumente usadas, descritas por Rossi, Lipsey e Freeman (2003, p. 54) como:

a. *Needs assessment*: questões sobre quais condições sociais um programa se destina a melhorar e a necessidade de formulação e implementação de um programa. Pode ser utilizada para avaliar um novo projeto ou um projeto já estabelecido, geralmente

fornece informações sobre quais serviços são necessários e como devem ser melhorados.

b. *Assessment of program theory*: trata-se da avaliação do contexto e do desenho do programa. É um plano que consiste essencialmente de suposições e expectativas, visando metas e objetivos. Geralmente é utilizada por agências de financiamento ou outros *decisionmakers* quando desejam lançar um novo programa.
c. *Assessment of program process (or process evaluation)*: é a avaliação do processo ou da implementação. Analisa questões relativas à fidelidade e efetividade das operações do programa, implementação e fornecimento dos serviços. Permite a checagem de metas e ajuda a institucionalizar formas de gerenciar sistema de informação e rotinas, aumentando o desempenho gerencial.
d. *Impact assessment*: preocupa-se com os resultados e os impactos de um programa. Avalia se os resultados foram alcançados e se houve efeitos não intencionais. Trata de isolar os efeitos exógenos que poderiam influenciar na análise dos resultados de um programa, por meio da estimativa de *net effects* (efeitos líquidos) de um programa. Essa metodologia é geralmente recomendada para programas estabelecidos há mais tempo, cujo programa teórico é bem definido, dado o esforço necessário na consecução de tais avaliações.
e. *Efficiency assessement*: avalia o custo do programa e custo-efetividade. Tem forte apelo comparativo face a desenhos alternativos, uma vez que muitos programas efetivos podem não ser atraentes por causa dos altos custos relativos a programas similares. Divide-se em *cost benefit analysis* (análise custo-benefício) e *cost-effectiveness analysis* (análise custo-efetividade). O custo-benefício relaciona os custos de empreender um programa com os benefícios gerados pelo programa, transformando-os em unidade de valor, geralmente unidade monetária. O custo-efetividade é expresso em termos de custos para se atingir um dado resultado, e a eficiência de um programa será dada em relação ao valor monetário para cada dada unidade de resultado.

Essa tipologia é amplamente utilizada no campo teórico das avaliações, e corroborada por outros teóricos. Cano (2006, p. 100) afirma que:

> [...] Normalmente, as avaliações incluem dois componentes: avaliação de processo ou de implementação e avaliação de impacto ou de resultado. A primeira tenta esclarecer em que medida o programa foi implementado conforme o plano original. A segunda, a mais importante, procura verificar se os efeitos finais foram atingidos [...] Algumas avaliações têm como objetivo específico calcular os custos do programa em relação ao impacto produzido. Assim, temos a avaliação de custo-benefício, que visa apurar o benefício monetário do programa para cada unidade de custo nele investida [...] existe também a avaliação de custo-efetividade, cujo objetivo é determinar o custo monetário por unidade de melhora produzida pelo programa.

Além disso, o autor compreende que as avaliações podem cumprir o papel de apenas informar dados (modelo mínimo de avaliação) ou pode, além disso, formular juízos de valor; tentar averiguar a razão do sucesso ou do fracasso do programa; se preocupar com o uso da própria avaliação; medir não apenas as variáveis dependentes que a intervenção pretende melhorar, mas qualquer outra capaz de ser diretamente influenciada; dentre outros Cano (2006, p. 99-104).

Salutar é perceber que, embora as avaliações cumpram funções diferentes em torno dos seus objetivos (avaliar processos, resultados, custos, eficiência, etc.), não se pode negar a natureza de dependência entre elas. Uma avaliação de impacto, por exemplo, poderá obter resultados pífios, cujos indicadores não apresentem ganho líquido aos participantes de um programa, indicando que o programa em tela deva ser descontinuado. Mas tais resultados podem advir de uma má implementação durante o processo de execução da política.

Um programa bem desenhado para alfabetizar jovens e adultos, com objetivos claros e metodologia pedagógica adequada, pode fracassar porque o transporte para levar os *policytakers* para os encontros presenciais falhava muitas vezes, implicando em duas consequências graves: evasão e baixo desempenho daqueles que se mantiveram no programa até o fim.

Nesse sentido, elaborar avaliações de impacto *(impact assessment)* sem levar em consideração o processo *(process evaluation)* poderá gerar grandes vieses aos resultados encontrados, cujo juízo de valor sobre a política implementada conduzirá a indicativos erráticos acerca da manutenção ou extinção do programa.

Escolhido o tipo de avaliação cabe a seguinte questão: quem a conduzirá? Existem várias maneiras de se aplicar uma avaliação, tais como avaliação externa e interna, ou uma combinação entre elas, avaliação mista, e, ainda, a autoavaliação.

A avaliação externa caracteriza-se pela inserção de pessoas que não fazem parte do órgão ou instituição executora do programa, por meio da contratação de avaliadores experientes. Na interna, colaboradores da instituição, não executores do programa, são escolhidos para o processo de avaliação, tentando realizá-lo com o máximo de neutralidade possível. A mista destina-se a programas em que colaboradores da instituição são orientados por técnicos externos à organização. E na autoavaliação os próprios responsáveis pela execução avaliam seus trabalhos, indicando o atingimento de metas e objetivos, no sentido de determinar o impacto ou efeito da política, ou simplesmente verificar o *quantum* do proposto foi cumprido.

Cada forma de aplicar a avaliação possui vantagens e desvantagens. A esse respeito, pode-se discorrer que:

> [...] Quando se realiza a avaliação interna, principalmente em órgãos públicos, corre-se o risco de os responsáveis ressaltarem aspectos bons ou positivos e minimizar os aspectos negativos. Em outros casos, procura-se o responsável pelo fracasso, quer seja interno ou externo à instituição. As vantagens da avaliação interna se ancoram na familiaridade com o objeto a ser avaliado, porém pode haver riscos provenientes da subjetividade, pois há menor objetividade em consequência do envolvimento dos avaliadores com o que se avalia, por ser ao mesmo tempo juiz e parte interessada. Os avaliadores alheios à organização apresentam maior objetividade, porém há dificuldade de captar plenamente todos os fatores em jogo de acordo com a natureza e funcionamento do programa. Assim, defende-se uma avaliação mista que busque equilibrar os fatores desfavoráveis ou favoráveis ao programa (WORTHEN; SANDERS; FITZPATRICK, 2004; AGUILAR; ANDER-EGG, 1994, apud MENDES et al., 2013, p. 100).

Na educação brasileira, no âmbito do ensino fundamental, têm-se uma questão interessante nesse particular. O principal índice nacional para o setor, o IDEB, utiliza-se de avaliação externa, sem participação dos *stakeholders* locais, seja na elaboração dos testes aplicados aos alunos, seja no desenho da pesquisa, ou nas suas dimensões. Uma avaliação desse tipo desconsidera uma série de valorações importantes, dentre elas, os aspectos locais e o contexto sociodemográfico regional; a contribuição dos *stakeholders* na formulação de objetivos e propósitos da avaliação e dos testes; o tipo de organização dos sistemas municipais de ensino, se seriados ou, de forma alternativa, utilizam-se de modelos pedagógicos baseados em ciclos de formação humana ou outros; e o debate pós-avaliativo, fundamental para a correção de rumos e consolidação de propostas junto aos *stakeholders*.

Faz-se necessária uma última reflexão nessa seção, referente à forma como a avaliação e os avaliadores tratam os *stakeholders*, e não somente eles, mas também todo o material e as informações coletadas por meio da metodologia utilizada durante o processo de avaliação. Rossi, Lipsey e Freeman (2003, p. 405) citam alguns princípios constantes no Guia de Princípios ao Avaliador (*Guiding Principles for Evaluators – American Evaluation Association)*:

> *[...] Integrity/honesty: evaluators ensure the honesty and integrity of the entire evaluation process. [...] Respect for people: evaluators respect the security, dignity, and self-worth of the respondentes, program participants, clientes, and other stakeholders with whom they Interact. [...] Responsibilities for general and public welfare: evaluators articule and take into account the diversity or interests and values that may be related to the general and public welfare.*

No entanto, tais princípios possuem relação direta com o avaliador e sua prática, mas existem outras questões éticas que se referem aos meios e modelos utilizados pela avaliação na tentativa de determinar os efeitos de uma política.

3.1.3 Avaliação de programas e avaliação de sistemas

Um programa recém-lançado, a depender do objeto que desejar intervir, pode demorar dias, meses ou anos para apresentar algum resultado. Uma lei regulamentada que proíbe qualquer cidadão de jogar lixo ou qualquer tipo de dejeto na rua tem efeitos logo após a sua publicação, porém mesmo que os termos da lei sejam adequados, seu *enforcement* pode ser deveras complicado, culminando em anos de implementação com resultados que se apresentam de forma perene, ou até mesmo não se efetivem.

Em contrapartida, políticas ainda mais complexas, como o Plano Nacional de Educação do Brasil, impõem dificuldades ainda maiores. A meta 5 desse Plano (MEC/SASE, 2014, p. 10) objetiva "[...] alfabetizar todas as crianças, no máximo, até o final do 3º (terceiro) ano do ensino fundamental". Para avaliar os efeitos de uma medida como essa, requer considerar um número significativo de influências nos sistemas municipais de ensino (responsáveis pelo nível de ensino fundamental) que levariam a tais resultados, uma vez que existem aspectos administrativos, pedagógicos e financeiros envolvidos. Há que se considerar questões relacionadas à formação e valorização do professor, alimentação escolar, material didático, equipamentos escolares, modelos de gestão das secretarias municipais, gestão das escolas, dentre outros tantos fatores. Neste caso, como atribuir resultados da alfabetização em três anos a apenas um único programa[12], dada a diversidade de variáveis exógenas ao seu desenho?

Avaliar um programa novo ou já estabelecido e avaliar sistemas de políticas públicas em educação, saúde, segurança pública, assistência social dentre outros, embora reservem entre si conceitos e técnicas similares, requer cuidados específicos na formulação da *program theory*. Isso em razão não somente dos objetivos desses dois momentos de avaliação, mas, sobretudo, dada a dimensão e a complexidade às quais tais análises seriam submetidas.

Na educação, por exemplo, pode-se avaliar um programa de reforma de escolas definido para o período de quatro anos, por meio dos ganhos líquidos implicados na segurança e no conforto da comunidade escolar a partir das intervenções, mas não haveria como saber o impacto da melhoria infraestrutural em um dos objetos mais ao centro da educação: fazer com que os alunos aprendam e se desenvolvam, contribuindo com o bem comum.

Este último objetivo, entretanto, divide-se em duas questões diferentes. Primeiro, o aluno aprende e se desenvolve. A concretização desse propósito, *per se*, gera um problema estrutural significativo, com diversas ameaças à avaliação, pois em um dado país, Estado, região, ou município, desenvolvem-se infindáveis projetos. Saber quais dos projetos representam impactam nesse objeto, dimensionando e mensurando seus resultados com grande validade e confiabilidade, é muito difícil, quiçá, impossível de ser realizado. Não restam muitas opções em uma avaliação desse tipo, a não ser definir dimensões importantes do setor educacional e tentar estabelecer relações lógicas entre elas, no sentido de compreender quais aspectos interferem mais em tal objeto.

Segundo, avaliar como o desenvolvimento das pessoas, advindo da escola, interfere e contribui com o bem comum é ainda mais intangível. Uma avaliação desse tipo, além de extrapolar os efeitos da educação formal escolar, envolve questões subjetivas de alto valor, como por exemplo a educação doméstica ou as condições sociais. É muito mais provável que qualquer avaliação de sistemas, visando sua validade interna, defina metas acerca da primeira questão e não da segunda.

O debate em torno dessas questões aprofunda-se na observação de Franco, Brooke e Alves (2008, p. 626-627) ao constatarem que:

> Especialmente para a educação básica, em âmbito nacional há o Sistema de Avaliação da Educação Básica que, atualmente é composto pelo SAEB e pela Prova Brasil. Destacam-se também as diversas experiências estaduais e municipais de avaliação. Todos esses projetos de avaliação têm o mérito de monitorar a situação educacional mediante uma seqüência de exercícios que, com métrica comparável ao longo do tempo, apresentam as tendências na evolução da qualidade da educação brasileira. [...] No entanto, as mencionadas experiências de avaliação da educação não oferecem os dados mais adequados para inferências causais acerca de quais políticas e práticas fazem diferença em educação. [...] enquanto uma seqüência de resultados do SAEB oferece boa orientação sobre a tendência da qualidade da educação nacional, ela não oferece a possibilidade de estudo pormenorizado sobre quais fatores promovem ou conspiram contra a qualidade.

Tais estudos são *surveys* seccionais, ou seja, são compostas muitas vezes apenas de varáveis-estoque, em dado momento do tempo mensuram o desempenho dos estudantes e fazem levantamento situacional das escolas, fazendo com que existam limitações importantes a serem consideradas:

---

[12] Existem vários programas que tentar melhoras as condições de alfabetização das crianças no Brasil. O mais conhecido deles é o Pacto Nacional pela Alfabetização em Idade Certa, desenvolvido pelo Governo Federal em colaboração com os Estados, o Distrito Federal e os Municípios. Mas também são executados diversos intentos pelas próprias Secretarias Municipais de Ensino em todo o território nacional.

> [...] a medida de desempenho em leitura ou em matemática é um agregado do aprendizado dos alunos ao longo de muitos anos. Já as medidas escolares disponíveis referem-se às condições escolares no ano da coleta de dados. Esta falta de sintonia temporal entre a medida do desempenho e as medidas das condições escolares, fragiliza as análises e inviabilizam a formulação de políticas de qualidade e equidade baseadas em evidências sólidas (FRANCO, BROOKE E ALVES 2008, p. 627).

O IDEB, por exemplo, trata-se de um indicador de qualidade educacional que propõe a combinação de informações de desempenho em exames padronizados pelo INEP aplicado aos estudantes ao final das etapas de ensino fundamental (5ª e 9ª séries do ensino fundamental e 3ª série do ensino médio) com informações sobre rendimento escolar, ou seja, a aprovação.

Um indicador desse tipo, embora tenha validade interna suficiente para mensurar resultados a que se propõe, não explica a casualidade entre os diversos aspectos de um sistema de ensino, pois se limita a verificar os condicionantes relativos aos resultados dos testes de proficiência e reprovação, ou aprovação escolar. Tal dificuldade é reconhecida por Fernandes (2007, p. 16-17) ao afirmar que:

> [...] vários aprimoramentos são possíveis, como, por exemplo, incluir a dispersão das notas, ao invés de se considerar apenas o desempenho médio. Por outro lado, seria necessário aprimorar nosso entendimento de como as escolas podem afetar o desempenho médio dos concluintes; isso nos permitiria adotar uma escolha mais criteriosa da forma funcional do Ideb. Por fim, e mais importante, seria necessário avançar nossos conhecimentos sobre as consequências, para a vida futura dos estudantes, de se adotar diferentes padrões de aprovação por parte das escolas, o que nos permitiria produzir um indicador cujo objetivo fosse o de maximizar o "bem-estar" dos alunos.

De maneira muito semelhante, são o Sistema Mineiro de Avaliação da Educação Pública – Simave, o Índice de Desenvolvimento da Educação do Estado de São Paulo – IDESP, e o Índice de Desenvolvimento do Estado do Rio de Janeiro – IDERJ[13], todos experiências estaduais de avaliação de sistemas. Já o Programa de Avaliação da Educação Básica do Espírito Santo – Paebes, avança um pouco mais na determinação de casualidades, uma vez que além do desempenho escolar averigua o "[...] clima escolar por meio de questionários contextuais vinculados à avaliação", conforme afirma a Secretaria da Educação do Estado do Espírito Santo (SEDU, 2014).

Estudo ainda mais abrangente foi realizado por meio de uma iniciativa que reuniu uma série de centros universitários parceiros, na tentativa de verificar a eficácia escolar em cinco grandes cidades brasileiras. O projeto, denominado Geres, possuía desenho longitudinal[14], realizado por um período de quatro anos. Diferente das avaliações anteriores, tal iniciativa faz uma análise multinível considerando a aprendizagem dos alunos e diversas variáveis sociodemográficas relativas aos próprios alunos, aos pais dos alunos, aos professores e gestores, e à infraestrutura escolar, possibilitando um universo muito maior de informações. (BROOKE e BONAMINO, 2011, p. 49)

Essa questão está presente em Raudenbush e Willms (1991, p. 2) quando discutem a unidade de análise da pesquisa em educação:

> [...] *should the researcher use the student as the unit, ignoring the differing organisational contexts in which education occurs and is reformed? Or should the organisational context (e.g., the classroom or school) be the unit of analysis, requiring the aggregation of student data within each context, ignoring all variation among students within a given context?*

Para os autores, muitos pesquisadores acreditam que o sucesso ou o fracasso de programas de reforma na área educacional dependem significativamente do contexto político-social, das particularidades das organizações e comunidades, e da complexa rede de interações entre estudantes, pais, professores, e administradores de cada nível do sistema (RAUDENBUSH E WILLMS, 1991, p. 4).

Em um trabalho sobre análise multinível em educação, na discussão relativa a pesquisas que procuram identificar efeitos provocados pelo ensino (*school effects research*), Lee (2001, p. 68) também corrobora com a necessidade de ampliar o nível de captação das variáveis, pois afirma que recentes artigos mostram que avaliações em educação têm frequentemente chegado a conclusões incorretas, por usarem procedimentos metodológicos

[13] A respeito desses programas ver Silva (2007), SEE-SP (2013) e SEDU (2012).

[14] Nesse tipo de desenho, assume-se "[...] a perspectiva de um corte longitudinal para que os mesmos alunos sejam observados ao longo do processo de escolarização." (BROOKE e BONAMINO, 2011, p. 17)

inadequados. Geralmente, os efeitos do contexto em que se desenvolve a pesquisa são subestimados em análises de um único nível.

Estudos longitudinais que levam em consideração modelos de análise multinível podem captar os efeitos das intervenções realizadas pelas instituições educacionais em relação às condições de aprendizagens pré-existentes dos estudantes, por meio de inferências estatísticas válidas, podem também ser replicadas ao longo do tempo e do espaço, tendo como resultado uma explicação plausível para processo pelo qual as escolas tornam-se eficazes. (GOLDSTEIN, 1997, p. 376)

Técnicas estatísticas educacionais que envolvem análises simples, levando em consideração apenas um nível de determinada unidade, a exemplo do aluno, tendem a desconsiderar o fato de que alunos fazem parte de classes, e estas partes da escola. Desconsiderar esse contexto deve levar a incoerências no todo do resultado, uma vez que micro e macro informações tomam parte de um mesmo modelo que não possui alcance técnico para dar conta dos efeitos de uma política. Leeuw e Meijer (2008, p. 1) acreditam que "*have predictors for variables of all these levels, and the challenge is to combine all these predictors into an appropriate statistical analysis, more specifically a regression analysis*".

Os autores acentuam que muitas técnicas podem ser usadas em análises multinível, a exemplo de análise multivariada, mas admitem que alguns cuidados sejam necessários, face ao que observam:

> *There is a clever way, used by Goldstein [46, Chapter 6], to fit general multivariate data into the multilevel framework. If we have n observations on m variables, we can think of these m observations as nested in n groups with m group members each. This amounts to thinking of the n×m data matrix as a long vector with nm elements and then building the model with the usual regression components and a suitable specification for the dispersion of the within-group disturbances. It is quite easy to incorporate missing data into this framework, because having datamissing simplymeans having fewer observations in some of the groups. On the other hand, in standard multilevel models, parameters such as regression coefficients are the same for different observations within the same group, whereas in multivariate analysis, this is rarely the case. Thus, writing the latter as a multilevel model requires some care (*LEEUW E MEIJER, 2008, p. 1).

Nesse sentido, a análise multivariada é uma boa técnica para a formulação de indicadores que promovam análise multinível, contando com algumas dimensões analíticas e com muitas variáveis aninhadas dentro delas, abordando os diversos contextos da educação escolar.

3.2 Experimentos aleatórios e não-aleatórios, validades e ameaças à pesquisa

Esta seção dedica-se ao tratamento conceitual de questões que compõem diretamente uma pesquisa ou programa de avaliação, que sem tais observâncias, um desenho poderá sucumbir perante às técnicas utilizadas e aos meios pelos quais obtiveram-se os resultados.

3.2.1 Definição de experimento

A ideia de experimento nas ciências físicas e naturais é diretamente relacionada à eliminação de interferências nos estudos, por isso são conduzidos em padrões adequados dentro de laboratórios montados especialmente com a finalidade de introduzir uma suposta causa e se verificar o correspondente efeito (CANO, 2006, p. 19). Já nas ciências sociais, o laboratório é a própria sociedade, instituição, organização, dentre outros, em que se pesquisam alguns resultados relativos a uma política pública.

Ao se desenhar um experimento, frequentemente supõe-se um modelo para descrever os dados a serem coletados e encontra-se um desenho eficiente que satisfaça uma ou mais condições dentro de uma pesquisa. Tal procedimento funciona bem quando se tem total clareza de que a técnica assumida dará conta de explicar as causas de um problema com os dados existentes, no entanto, raramente isso acontece. Por isso, o teste de diversos modelos é fundamental para tratar qualquer experimento (GHOSH, 2012, p. 283).

Corroborando a respeito, Shadish, Cook e Campbell (2002, p. 12) definem experimento como "[...] *a study in which an intervention is deliberately introduced to observe its effects*". O efeito pretendido será a diferença entre o que de fato aconteceu e o que deveria ter acontecido, por isso o atributo comum em todos os experimentos é o controle do tratamento.

Os autores apresentam quatro tipos de descrições relativas a experimentos, alertando para o fato de pesquisas que não podem ser consideradas experimentos por sua própria natureza investigativa:

> ***Randomized Experiment*** [grifo nosso]*: an experiment in which units are assigned to receive the treatment or an alternative condition by a random process such as the toss of a coin or a table of random numbers.*
> ***Quasi-Experiment*** [grifo nosso]*: An experiment in which units are not assigned to conditions randomly.*
> ***Natural Experiment*** [grifo nosso]*: Not really an experiment because the cause usually cannot be manipulated; a study that contrasts a naturally occurring event such as an earthquake with a comparison condition.*
> ***Correlational Study*** [grifo nosso]*: Usually synonymous with nonexperimental or observational study; a study that simply observes the size and direction of a relationship among variables (*SHADISH, COOK E CAMPBELL, 2002, p. 13-18*).*

A escolha de cada experimento levará a um tratamento diferente concernente às técnicas a serem usadas ao longo de uma avaliação, de forma que uma ou mais variáveis independentes serão manipuladas para observar seus efeitos em uma ou mais variáveis dependentes (YAREMKO et al., 1986, p. 72).

Os principais tipos de pesquisa, entretanto, mais utilizados na avaliação de políticas sociais são: *Randomized Experiment* (experimentos aleatórios ou randômicos) e *Quasi-Experiment* (quase-experimento ou experimentos não aleatórios).

Cano (2006, p. 20) afirma que para que uma pesquisa seja considerada um experimento aleatório nas ciências sociais, levam-se em conta três requisitos básicos: a introdução da causa[15] é feita pelo pesquisador, não acontecendo naturalmente, mas artificialmente, não restando dúvidas sobre qual seja a causa e qual o efeito; o controle da situação experimental, em que é exercido controle sobre o processo e o contexto, buscando mensurar apenas os efeitos da variável introduzida, eliminando influências não procedentes ao programa; e a designação aleatória dos participantes do programa (pessoas, escolas, organizações, etc.) para grupos experimental e de controle[16], pois a lógica experimental parte de um contrafactual impossível.

O contrafactual neste caso pode ser exemplificado da seguinte forma: um programa é implementado visando à alfabetização de crianças em dado sistema de ensino, por meio de formação de professores, material didático adequado e ampliação dos turnos de aula. Verificar se o programa deu certo significa mensurar os efeitos do programa na leitura, na escrita e no letramento de crianças submetidas ao programa antes e depois da intervenção.

Ocorre que todas as crianças aprenderiam independentemente de estarem submetidas às ações da política pública, por diversos fatores, sejam familiares, escolares, ou outros. Então, não seria adequado promover uma avaliação que pretendesse chegar aos resultados de um programa apenas pela verificação do que essas sabiam antes e depois de participarem das atividades, mas é preciso observar também o que ocorreu com crianças não submetidas ao programa, que possuam características muito semelhantes às do grupo de intervenção, para a partir daí, passar a fazer comparativos e mensurar o ganho líquido do programa.

O supracitado contrafactual está exatamente na necessidade do grupo de controle, uma vez que uma criança não pode estar participando, ao mesmo tempo, dos grupos de controle e de intervenção, pois os efeitos da alfabetização são permanentes. Dito de outra forma, "[...] não faria sentido voltar a medir a pessoa num segundo momento, depois de retirar a causa. [...] como se trata de dois momentos diferentes, existe a possibilidade de que alguma outra coisa aconteça nesse intervalo que mude a medição" (CANO, 2006, p. 20).

Depreende-se daí que existem limitações de toda ordem (financeiras, administrativas, logísticas, ambientais, éticas) ao uso de experimentos em ciências sociais, motivo pelo qual tendências teóricas passaram a demandar outras formas de averiguar causas e efeitos nessa área.

---

[15] É o conjunto de condições sem as quais não haveria determinado resultado, ou ainda é o que causa o efeito. Para Mackie (1974) apud Shadish, Cook e Campbell (2002, p. 4) a causa, ou *inus condition*, é "*an insufficient but nonredundant part of an unnecessary but sufficient condition*". É insuficiente pois existem outros fatores, ou causas, que podem explicar um dado fenômeno. É não-redundante somente se os efeitos dependerem de questões contextuais do ambiente em análise.

[16] As técnicas utilizadas por meio de várias pesquisas em ciências sociais levam em consideração dois grupos compostos por unidades semelhantes. Um deles, o de intervenção ou experimento, recebe os serviços do programa, enquanto o outro, o de controle, não participa da intervenção. Os ganhos líquidos, caso existam, serão observáveis a partir da diferença entre eles.

Questões éticas podem ser visualizadas, por exemplo, em setores públicos como a educação. Como realizar experimentos para saber se o uso adequado de tecnologia da informação pode melhorar a cognição de adolescentes de dado sistema municipal de ensino, nas diversas áreas do conhecimento? Um dos desenhos possíveis, para simplificar o caso, seria criar dois grupos aleatórios de estudantes (grupos de controle e de intervenção), com características semelhantes e chegar às conclusões após intervenção.

Entretanto, levando apenas em conta a questão da aprendizagem, sem imbricações ligadas ao desenvolvimento do conhecimento, o que conduziria a uma ampla discussão teórico-conceitual, seria necessário um tempo suficiente para que os efeitos líquidos das intervenções fossem mensurados nas áreas de matemática, expressão, geografia, história, dentre outras. Durante esse período, de forma aleatória, uma série de estudantes estariam privados das atividades relacionadas à política pública. Estatisticamente, poder-se-ia justificar a necessidade desse empreendimento, mas como fazê-lo socialmente?

Um frequente obstáculo ao uso de experimentos aleatórios decorre do fato de que muitos *stakeholders* possuem escrúpulos éticos relativos à aleatorização, vendo-as como arbitrárias, sobretudo por privar o grupo de controle de benefícios positivos. O contra-argumento é óbvio, ordinariamente, não se conhece se uma intervenção será efetiva, daí a necessidade de um experimento. (ROSSI, LIPSEY e FREEMAN, 2003, p. 259)

Tais limitações à pesquisa aleatória levaram aos quase-experimentos, que possuem rigor técnico, mas não atendem a todos os requisitos experimentais[17].

Esse tipo de pesquisa tem o mesmo propósito dos outros experimentos que é testar hipóteses causais descritivas sobre causas manipuláveis, usando grupos de controle e, ou testes realizados a *priori*, no sentido da garantia de boa inferência contrafactual. A grande diferença é que em experimentos dessa natureza, falta atribuição aleatória às unidades de análise. Então, as condições relativas a uma amostra dar-se-ão "*by means of self- selection, by which units choose treatment for themselves, or by means of administrator selection, by which teachers, bureaucrats, legislators, therapists, physicians, or others decide which persons should get which treatment*" (SHADISH, COOK e CAMPBELL, 2002, p. 14).

No entanto, conforme asseguram Campbell e Stanley (1963, p. 34-35):

> *There are many natural social settings in which the research person can introduce something like experimental design into his scheduling of data collection procedures (e.g., the when and to whom of measurement), even though he lacks the full control over the scheduling of experimental stimuli (the when and to whom of exposure and the ability to randomize exposures) which makes a true experiment possible. Collectively, such situations can be regarded as quasi-experimental designs*.

De tal questão, deriva-se que a introdução de pesquisas quase-experimentais melhorou a *evaluability*[18] (avaliabilidade) de muitos programas, levando os pesquisadores a terem considerável controle sobre suas mensurações, sobre como a atribuição não-aleatória é executada e sobre os tipos de grupos de comparação com os quais os grupos de tratamento serão comparados. Sendo assim, os desenhos quase-experimentais tendem a ser muito mais flexíveis, o que permite sua aplicação em inúmeros casos em que os experimentos não seriam possíveis (CANO, 2006, p. 25, 69; SHADISH, COOK e CAMPBELL, 2002, p. 14).

A disseminação de pesquisas desse tipo propiciaram a criação de diversos desenhos quase-experimentais alternativos, concebidos por meio de pré-testes e pós-testes aplicados a uma amostra não aleatória, desenhos com grupos de controle não equivalentes, dentre outras tantas variações[19].

Além disso, a escolha do desenho e seu programa teórico devem considerar algumas condições no sentido de garantir sua validade (interna ou externa), reduzindo e isolando todas as possíveis ameaças ao modelo avaliativo.

A validade de um dado programa refere-se ao nível de confiança que um conjunto de informações, advindas de inferências causais, representam, de fato, resultados de políticas públicas sobre o objeto-problema gerador da política pública. Então constatar a validade de um *program theory* é "*make a judgment about the extent to which relevant evidence supports that inference as being true or correct*" (SHADISH, COOK e CAMPBELL, 2002, p. 34).

---

[17] O controle do contexto e das variáveis intervenientes é muito pequeno; existem grupos de controle e experimental, mas suas designações não são realizadas de modo aleatório, não sendo possível garantir a equivalência de ambos (CANO, 2006, p. 25);

[18] "[...] indica a capacidade de uma intervenção de ser efetivamente avaliada. Alguns programas são dificilmente avaliáveis e isso deve ser considerado antes de decidir realizar uma avaliação". (CANO, 2006, p. 100)

[19] A esse respeito, ver Campbell e Stanley (1963); Cano (2006, cap. 7) e Rossi, Lipsey e Freeman (2003, cap. 9).

Os autores chamam a atenção, no entanto, para o fato de que os julgamentos com base em validades nunca são absolutos, por existirem vários graus de validade que podem ser invocados em uma inferência causal, levando sempre à compreensão de que o resultado de um dado teste possui é, aproximadamente, ou provisoriamente, verdadeiro ou falso, válido ou inválido, pela seguinte razão:

> *Validity is a property of inferences. It is not a property of designs or methods, for the same design may contribute to more or less valid inferences under different circumstances. For example, using a randomized experiment does not guarantee that one will make a valid inference about the existence of a descriptive causal relationship. After all, differential attrition may vitiate randomization, power may be too low to detect the effect, improper statistics may be used to analyze the data, and sampling error might even lead us to misestimate the direction of the effect. So it is wrong to say that a randomized experiment is internally valid or has internal validity-although we may occasionally speak that way for convenience. The same criticism is, of course, true of any other method used in science, from the case study to the random sample survey. No method guarantees the validity of an inference* (SHADISH, COOK e CAMPBELL, 2002, p. 34).

Existem vários tipos de validade, e seus conceitos foram aperfeiçoados ao longo dos anos, desde o importante trabalho de Donald Campbell[20], em 1957, até os dias atuais. Tais corroborações ao conceito podem ser observadas na divisão teórica entre os seguintes tipos de validade: *Statistical Conclusion Validity* (validade da conclusão estatística)*, Internal Validity* (validade interna), *Construct Validity* (validade de construto), *External Validity* (validade externa) (SHADISH, COOK e CAMPBELL, 2002, p. 38).

A primeira implica em aferir qual é a inferência estatística sobre a correlação ou covariância[21] entre o tratamento aplicado aos *targets* e os resultados do programa. A validade interna, por sua vez, visa verificar a validade das inferências sobre se a correlação observada entre o tratamento (A) e os resultados (B) refletem uma relação causal de A e B com as variáveis que foram manipuladas e mensuradas pela pesquisa. Cano (2006, p. 29) define esse tipo de validade como:

> [...] o grau de certeza de que o efeito na variável dependente do experimento foi causado pela variável independente do experimento. Em outras palavras, é a confiança de que foi a causa pesquisada, e não outro fator, que produziu os efeitos observados [...] a validade interna faz referência à inferência causal entre causa e efeito tal como foram definidos e operacionalizados no experimento, sem pretensão de generalização.

A validade de construto refere-se àquela que tem por objetivo inferir sobre os construtos de ordem superior, que representam elementos da amostragem, ou seja, é o grau em que as inferências são a garantia de que as pessoas observadas, as definições adotadas na pesquisa e a relação causa-efeito são significativas dentro da pesquisa.

A validade externa é a inferência sobre se a relação causa-efeito encontrada em uma dada avaliação pode ser extrapolada a mais pessoas, contextos, variáveis tratadas e mensuradas. Nessa questão, imbricam-se tanto os *targets* e o desenho de pesquisa, quanto os tratamentos dados às variáveis e às observações.

Cano (2006, p. 21) observa que um *program theory* deve considerar a validade interna como condição essencial à pesquisa, pois de nada adiantaria tentar extrapolar determinados resultados a outros contextos sem que se tenha clareza da fidedignidade dos resultados estatísticos observados pela pesquisa. Sugere ainda que “o ideal é atingir um alto grau dos dois tipos de validade, mas amiúde é necessário um equilíbrio entre ambos, especialmente nos casos em que a perseguição de um tende a diminuir o outro”.

Tal perseguição ocorre em razão de a validade interna ser tanto mais forte quanto maior for o controle de dado experimento sobre as variáveis envolvidas, enquanto a extrapolação de determinados resultados para um todo (pessoas, processos, observações, contextos), por vezes só é possível com a inserção de novas variáveis, mais difíceis de controlar.

Não obstante, existem diversas ameaças às validações de uma pesquisa. Shadish, Cook e Campbell (2002, p. 53-101) relacionam diversas delas para cada tipo de validade. A validade interna pode ser prejudicada caso, por exemplo, ocorram as seguintes situações: precedência temporal ambígua, ou seja, falta de clareza sobre o que se deu primeiro em um

[20] *Factors relevant to the validity of experiments in social settings*, 1957.
[21] É a medida do grau de interdependência entre duas variáveis aleatórias; verifica se existe alguma dependência entre elas e em que nível.

dado contexto, o que impossibilita estabelecer a causa; a seleção da amostra é feita entre grupos não-equivalentes; durante o tratamento ocorrem eventos exógenos que podem modificar os efeitos do programa, assim como naturalmente ocorrem mudanças ao longo do tempo que se refletem nos *targets* e podem ser confundidas com o efeito do programa, mas não se referem ao programa e sim à história e à maturação; regressão em direção da média; atrito ou mortalidade amostral pode ocorrer quando se mede o mesmo grupo de *targets* muitas vezes; a submissão dos targets por mais de uma vez ao mesmo tipo de teste pode influenciar nos resultados finais; dentre outras.

Cano (2006, p. 42) afirma que "a lista de ameaças à validade interna ou externa não deve ser considerada uma receita infalível para o sucesso da pesquisa. Visa tão-somente contribuir para a reflexão sobre como melhorar o desenho da pesquisa para garantir uma inferência causal forte e uma generalização ampla".

4. CONSIDERAÇÕES FINAIS

A revisão bibliográfica apresenta uma robusta discussão de diversos autores sobre metodologias de avaliação de políticas públicas, enfocando áreas sociais, especificamente na área educacional, localizando o debate no Brasil.

Embora a discussão sobre conceitos relacionados à política pública e seus ciclos seja relativamente nova, iniciada em meados do século passado, verificou-se um crescente interesse pelas metodologias de avaliação de programas, sobretudo nas áreas sociais.

As principais discussões teóricas na área concentram-se nos tipos e possibilidades de avaliação, levando-se em consideração a arena política, seus atores e objetivos precípuos observados no cerne do problema público, além da preocupação quanto ao timing e usos da avaliação. Entretanto, os estudos carecem de maior aprofundamento científico no que se refere às avaliações dos sistemas públicos, já desenvolvidas, empiricamente, em diversas partes do Brasil, a exemplo do que já ocorre na área educacional.

REFERÊNCIAS

ANDERSON, C. W. The place of principles in policy analysis. **American Political Science Review**, v. 73, n. 3, p. 711-723, 1979.

ARRETCHE, M. T. S. Tendências no estudo sobre avaliação. In: **Seminário Avaliação de políticas sociais: uma questão em debate**, São Paulo: PUC-SP, 1998, 10 p.

AXELSEN, M. et al. Examining the role of is audit in the public sector. In: **PACIS 2011 Proceedings,** Brisbane-Australia: University of Queensland, 2011, 15 p. Disponível em: <http://aisel.aisnet.org/pacis2011/23>. Acesso em: 20 out. 2015.

BARZELAY, M. Instituições centrais de auditoria e auditoria de desempenho: uma análise comparativa das estratégias organizacionais na OCDE. **Revista do Serviço Público**, v. 2, n. 53, p. 5-35, 2002.

BOSCHETTI, I. **Questões correntes no debate sobre metodologias de avaliação de políticas públicas**. DF: UnB, 2006. 17 p.

BROOKE, N.; BONAMINO, A. **GERES 2005: razões e resultados de uma pesquisa longitudinal sobre eficácia escolar**. Rio de Janeiro: Walprint Gráfica e Editora, 2011. 263 p.

CAMPBELL, D. T.; STANLEY, J. C. **Experimental and quasi-experiment al designs for research**. London: Houghton Mifflin Company, 1963. 84 p.

CANO, I. **Introdução à Avaliação de Programas Sociais**. 3. ed. Rio de Janeiro: Editora FGV, 2006. 120 p.

CUNHA, C. **Avaliação de Políticas Públicas e Programas Governamentais: tendências recentes e experiências no Brasil**. Washington D.C.: George Washington University, p. 1-41, 2006.

DERLIEN, H.-U. Una comparación internacional en la evaluación de las políticas públicas. **Revista do Serviço Público**, v. 52, n. 1, p. 105-123, 2001.

FARIA, C. A. P. DE. A política da avaliação de políticas públicas. **Revista Brasileira de Ciências Sociais**, v. 20, n. 59, p. 97-110, 2005.

FÉLIX, C. L. Auditoria de desempenho aplicada na avaliação da execução de metas orçamentárias do setor público. In: **Congresso Brasileiro de Contabilidade**. Anais... Rio

Grande do Sul: Gramado, 2008, 12 p. Disponível em: <http://www.ccontabeis.com.br/18cbc/413.pdf>. Acesso em: 10 out. 2014

FERNANDES, R. Índice de Desenvolvimento da Educação Básica (IDEB). Brasília-DF: INEP. **Série documental. Textos para discussão**, n. 26, 2007, 7 p.

FIRME, T. P.; LETICHEVSKY, A. C.; DANNEMANN, Â. C. Evaluation culture and evaluation polic y as guides to practice : reflections on the Brazilian experience. **Ensaio: avaliação de políticas públicas educacionais**. Rio de Janeiro: Cesgranrio, v. 17, n. 62, p. 169-179, 2009.

FRANCO, C.; BROOKE, N.; ALVES, F. Estudo longitudinal sobre qualidade e equidade no ensino fundamental brasileiro: GERES 2005. **Ensaio: Avaliação e Políticas Públicas em Educação**. Rio de Janeiro: Cesgranrio, v. 16, n. 61, p. 625-637, 2008.

GERTLER, P. J. et al. **Impact Evaluation in Practice**. Washington DC: The World Bank, 2010, 266 p.

GHOSH, S. Search linear model for identification and discrimination. In: **Design and analysis of experiments: special designs and applications**. Virginia: John Wiley & Sons, 2012, p. 555.

GOLDSTEIN, H. Methods in School Effectiveness Research. In: **School effectiveness and school improvement**. London: University of London, v. 8, n. 4, p. 369-395, 1997.

HENRY, G. T. How Modern Democracies Are Shaping Evaluation and the Emerging Challenges for Evaluation. **American Journal of Evaluation**, v. 22, n. 3, p. 419-429, 2000.

KETTNER, P.; MORONEY, R.; MARTIN, L. **Designing and managing programs: An effectiveness-based approach**. 3. ed. Los Angeles: SAGE Publications Ltd, 2012. 300 p.

KHANDKER, S.; KOOLWAL, G.; SAMAD, H. **Handbook on impact evaluation: quantitative methods and practices**. Washington-D.C.: The World Bank, 2010. 239 p.

LEE, V. E. Using multilevel methods to investigate research questions that involve nested data: examples form education. **Estudos em avaliação educacional**, n. 24, p. 46–68, jul.-dez. 2001.

LEEUW, J. DE; MEIJER, E. **Handbook of Multilevel Analysis**. New York-NY: Springer New York, 2008. 504 p.

MEC. Ministério da Educação. **Planejando a Próxima Década: conhecendo as 20 metas do Plano Nacional de Educação**. Brasília-DF: MEC/SASE, 2014a. 63 p.

MENDES, C.V. et. al. Metodologia de avaliação de implementação de programas e políticas públicas. **EccoS Revista Científica**, São Paulo, n. 30, p. 93-111, 2013.

RAUDENBUSH, S. W.; WILLMS, J. D. (Orgs.). **Schools, Classrooms, and Pupils: international studies of schooling form a multilevel perspective**. San Diego-Califórnia: Academic Press, Inc., 1991. 260 p.

RIDGE, J. B. **Evaluation techniques for difficult to measure programs: for education, nonprofit, grant funded, business and human service programs**. USA: Xilibris Corporation, 2010. 261 p.

ROSSI, P. H.; LIPSEY, M. W.; FREEMAN, H. E. **Evaluation: a systematic approach**. 7. ed. London: SAGE Publications, Ltd, 2003. 470 p.

SANTOS, A. R. **Monitoramento e avaliação de programas no setor público: a experiência do PPA do Governo Federal no período 2000–2011**. 2012. Monografia (Especialização em Serviço Público). Brasília-DF: Instituto Serzedello Corrêa do Tribunal de Contas da União, 30 nov. 2012. 69 p.

SCHNEIDER, A. L. Pesquisa avaliativa e melhoria da decisão política: evolução histórica e guia prático. In: HEIDEMANN, F. G.; SALM, J. F. (Orgs.). **Políticas Públicas e Desenvolvimento - Bases Epistemológicas e Modelos de Análise**. 2. ed. Brasília: Editora UNB, p. 311–338, 2010.

SECCHI, L. **Políticas públicas: conceitos, esquemas de análise, casos práticos**. 2. ed. São Paulo: Cengage Learning, 2013. 168 p.

SEDU, Secretaria de Educação do Estado do Espírito Santo. PAEBES – alfa 1ª onda. Disponível em: <http://www.paebesalfa1onda.caedufjf.net/avaliacao-educacional/a-avaliacao/>. Acesso em: 14 out. 2014.

SEE. Secretaria de Educação do Estado de São Paulo. **Índices educacionais**. 2013. Disponível em: < http:// http://www.educacao.sp.gov.br/indices-educacionais> Acesso em: 12 fev. 2015.

SEEDUC, Secretaria de Educação do Estado do Rio de Janeiro. **Educação: SAERJ/SAERJINHO/IDERJ**. 2012. Disponível em: <http://www.rj.gov.br/web/seeduc/exibeconteudo?article-id=843535>. Acesso em: 14 out. 2014.

SHADISH, W. R.; COOK, T. D.; CAMPBELL, D. T. **Experimental and Designs for Generalized Causal Inference**. Boston: Houghton Mifflin Company, 2002. 623 p.

SILVA, M. J. DE A. O Sistema Mineiro de Avaliação da Educação Pública: impactos na escola fundamental de Uberlândia. **Reice - Revista Eletônica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación**, v. 5, n. 2, p. 241-253, 2007.
TREVISAN, A. P. A.; BELLEN, H. V. H. M. V. Avaliação de políticas públicas: uma revisão teórica de um campo em construção. **Revista de Administração Pública**, v. 42, n. 3, p. 529-550, 2008.
UNICEF. A UNICEF Guide for Monitoring and Evaluation: Making a Difference? New York: UNICEF Org., 1990. 92 p.
YAREMKO, R. M. et al. **Handbook of research and quantitative methods in psychology for students and professionals**. Hillsdale-NJ: Erlbaum, 1986. 335 p.