GUIA PARA AVALIAR POLÍTICAS PÚBLICAS | Vol. 4

E quando a política está em andamento? Avaliação *ex post*!

Instituto Jones dos Santos Neves

GOVERNO DO ESTADO DO ESPÍRITO SANTO
Secretaria de Estado de Economia e Planejamento

Instituto Jones dos Santos Neves

Guia para Avaliar Políticas Públicas | volume 4. E quando a política está em andamento? Avaliação *ex post*!

Vitória, ES, 2018. 122 p.; il. tab.

1.Políticas Públicas. 2. Avaliação de Políticas Públicas. 3. Projetos. 4.Planejamento. 5. Economia. 6. Administração Pública. 7. Espírito Santo (Estado). I.Cappellazzo Arabage, Amanda. II. Portela Fernandes de Souza, André. III. Seidel, Fernanda. VI. Weber Costa, Gabriel. V. Gomes de Macêdo Lacerda, Gabriela. IV. da Motta Silveira Borges, Ligia. VII. Silva e Lima, Lycia. IV. Título.

Elaboração
Amanda Cappellazzo Arabage*
André Portela Fernandes de Souza*
Fernanda Seidel**
Gabriel Weber Costa*
Gabriela Gomes de Macêdo Lacerda
Ligia da Motta Silveira Borges
Lycia Silva e Lima*

Colaboração
Andrezza Rosalém Vieira
Juliana Camargo*
Matheus Mascioli Berlingeri*
Thais Peres Dietrich*

Revisão
Ana Carolina Giuberti
Kátia Cesconeto de Paula
Magnus William de Castro

Editoração
Arthur Ceruti Quintanilha
João Vitor André

Bibliotecário
Jair Rosário Filho

* Centro de Aprendizagem em Avaliação e Resultados para o Brasil e a África Lusófona (FGV EESP Clear).
** Pesquisador Bolsista FAPES.

Agradecimentos

A elaboração do Guia para Avaliar Políticas Públicas é fruto de um processo amplamente colaborativo. Diversas instituições, gestores e técnicos estiveram engajados desde a concepção até a edição do documento.

Agradecemos o apoio e as valiosas contribuições da Secretaria de Estado de Economia e Planejamento, atuando intensamente em todas as etapas. Fundamental também o trabalho em conjunto com as instituições que atuam no Núcleo de Monitoramento e Avaliação: Secretaria de Estado de Gestão e Recursos Humanos, Secretaria de Estado da Fazenda, Fundação de Amparo à Pesquisa e Inovação do Espírito Santo e Escola de Serviço Público do Espírito Santo.

À Secretaria de Estado de Direitos Humanos, por viabilizar o arranjo institucional para a produção técnica.

Estamos profundamente agradecidos aos gestores e técnicos que compartilharam suas experiências em formulação de políticas e em monitoramento e avaliação de políticas capixabas, auxiliando na construção de exemplos que permeiam o Guia. São eles: Andressa Buss Rocha (SEDU), Anselmo Dantas (SESA), Antônio Ricardo Freislebem da Rocha (IJSN), Carlos Augusto Gabriel de Souza (SESP), Eduardo Malini (SEDU) , Filipe Lube (SEGER), Frederico Nogueira (IJSN), João Marcos Chipolesch (SEAG), Jonas Lisboa (SEAG), Julierme Tosta (SEFAZ), Kettini Upp Calvi (SEP), Letícia Campos Souza (SECONT), Magnus William de Castro (IJSN), Marcos Franklin Sossai (SEAMA), Marcelo Belumat (BANDES), Márcio Bastos Madeiros (SEP), Patrick Tranjan, Roberto de Freitas Campos (SEFAZ), Sandra Mageski (SEP), Sandra Mara Pereira (IJSN), Tânia Mara de Araújo (SESA), Thiago de Carvalho Guadalupe (IJSN), Victor Nunes Toscano (SETADES) e Vinicius Cappeletti .

Em particular, agradecemos à equipe do Instituto Jones dos Santos Neves, com a participação direta do Escritório de Projetos na coordenação e na produção técnica do trabalho e às Coordenações pelas sugestões, *feedback* e orientações.

À equipe do Centro de Aprendizagem em Avaliação e Resultados para o Brasil e a África Lusófona (FGV EESP Clear), da Fundação Getulio Vargas, pelo auxílio não apenas no Guia, mas em todas as etapas de implementação do Sistema de Monitoramento e Avaliação de Políticas Públicas do Espírito Santo, em que compartilharam conhecimento técnico e experiência com muita generosidade.

E por fim, à Andrezza Rosalém, por liderar a modelagem do sistema capixaba de avaliação de políticas públicas e abrir os caminhos para a confecção dessa importante ferramenta, agora à disposição de todo o Governo do Estado do Espírito Santo.

Sumário

Apresentação

"E quando a política está em andamento? Avaliação *ex post*!" é um dos volumes que integram o Guia para Avaliar Políticas Públicas, material de referência do Sistema de Monitoramento e Avaliação de Políticas Públicas do Estado do Espírito Santo, criado pela Lei nº 10.744, de 05 de outubro de 2017, inspirado nas melhoras práticas nacionais e internacionais.

O objetivo deste volume é orientar e auxiliar a avaliação de políticas, programas e projetos já em andamento. Prevista como uma das linhas de avaliação na Lei nº 10.744/2018, esse tipo de avaliação é um valioso instrumento para compreender as razões do sucesso ou das falhas de uma política. Essas informações são fundamentais para o aprimoramento das políticas adotadas.

As avaliações *ex post* podem responder a diversas perguntas e as metodologias aplicadas dependerão de quais questões se tem interesse em responder. Essas questões podem envolver desenho, processos, resultados ou custos de uma política. Incentivamos todos os gestores a avaliar seus programas e projetos já implementados.

Duas premissas foram importantes na elaboração deste Guia. A primeira é de que ele tivesse um caráter prático e linguagem acessível, auxiliando profissionais de políticas públicas no desenho das avaliações. Para estudos mais aprofundados, recomendamos uma vasta literatura. A segunda, é de que este documento espelhasse a realidade do Governo do Estado do Espírito Santo, para que técnicos e gestores de fato se apropriem dele. Por isso, ele foi idealizado e elaborado a muitas mãos. Contamos com a participação de diversos órgãos em todas as etapas. Ainda, ao refletir a identidade do Governo, entendemos que este Guia é dinâmico: na medida em que avançarmos na cultura de monitoramento e avaliação no âmbito do Espírito Santo, futuras edições poderão aprimorar as recomendações e incorporar novos elementos.

Informações complementares sobre como estruturar uma proposta de monitoramento podem ser consultadas no volume "Como monitorar políticas públicas?"; e sobre outros tipos de avaliações, nos volumes "A política é nova? Avaliação *ex ante*!" e "Avaliação ao alcance de todos: análise executiva".

Esperamos que o Guia para Avaliar Políticas Públicas, enquanto ferramenta disponível a todos, contribua para melhorar a eficiência do gasto do público, a qualidade da gestão e os resultados alcançados pelas políticas públicas.blico, a qualidade da gestão e os resultados alcançados pelas políticas públicas.

Gabriela Gomes de Macêdo Lacerda
Diretora Presidente do
Instituto Jones dos Santos Neves

Prefácio
Melhores Resultados

O mundo em que vivemos está sempre em transformação, mas, nas últimas décadas, principalmente em decorrência da disseminação da internet e das novas tecnologias de comunicação e informação, a velocidade com que essas mudanças ocorrem tem sido cada vez maior. Tecnologias que estão, a cada dia, mais presentes em nossas vidas.

Na esfera de governo, neste mundo complexo e conectado, crescem demandas da sociedade por políticas públicas eficazes e eficientes. Uma sociedade que, favorecida pelos avanços tecnológicos, é capaz de acompanhar a gestão e dela cobrar mais e melhor desempenho, em seu benefício.

Especialmente nestes tempos de restrição de recursos financeiros, para que possamos produzir os melhores resultados para a sociedade, é fundamental que esses recursos sejam aplicados com sabedoria.

Pensando assim, o Governo do Estado do Espírito Santo vem trabalhando e oferecendo aos cidadãos maior velocidade e eficácia no desenvolvimento e aplicação de ações e projetos. Com seu Modelo de Gestão da Estratégia, aumentou a capacidade de executar entregas relevantes, de alto benefício e grande poder de transformação na sociedade capixaba.

Com o Sistema de Monitoramento e Avaliação de Políticas Públicas ligado ao Ciclo de Planejamento e Orçamento, o Governo amplia ainda mais sua capacidade de trabalhar com eficiência e eficácia, contribuindo para melhorar o padrão do gasto público.

O presente Guia para Avaliar Políticas Públicas que gestores e servidores têm agora em mãos é uma ferramenta útil para que o Espírito Santo continue sendo exemplo para o Brasil, com melhores práticas em gestão pública.

Regis Mattos Teixeira
Secretário de Estado de
Economia e Planejamento

Prefácio
Caminhos para a Avaliação

Este Guia para Avaliar Políticas Públicas oferece aos gestores e servidores públicos o estado da arte das melhores práticas de monitoramento e avaliação de políticas públicas. Ele é apresentado em linguagem acessível e exposto em passos necessários para a boa utilização do SiMAPP – Sistema de Monitoramento e Avaliação de Políticas Públicas do Estado do Espírito Santo.

O guia está dividido em quatro volumes. O volume 1, intitulado "A política é nova? Avaliação *ex ante*", apresenta as etapas necessárias para a elaboração de um programa novo ou a ser reformulado. Ele contém os instrumentos a serem utilizados para o seu bom desenho. Busca-se, de um lado, a consistência entre o desenho da política e o problema a ser solucionado e, de outro, a consistência entre as ações e os resultados a serem alcançados. Para tanto, este primeiro volume se inicia com o diagnóstico do problema, passa pela criação do Modelo Lógico e finaliza com a análise de custos e efetividades esperadas.

O volume 2, "Avaliação ao alcance de todos: análise executiva", traz os procedimentos para a avaliação rápida e de baixo custo de um programa já em execução. O objetivo da análise executiva é apresentar um diagnóstico inicial, mas geral, do desempenho da política pública em seus aspectos de desenho, processos e resultados. Espera-se, ao final, obter um julgamento informado sobre os pontos fortes e fracos da política, de modo a embasar recomendações sobre possíveis ajustes, inclusive indicações para avaliações mais aprofundadas.

O volume 3, "Como monitorar uma política pública?", apresenta as etapas e as ferramentas para o monitoramento de um programa. Inicia-se com os critérios para a seleção e criação de indicadores e coleta de dados, passando por definições de metas até sugestões para o bom uso das informações sistematicamente geradas. Essas evidências servem para identificar mudanças e tendências a fim de embasar o processo de tomadas de decisões.

Por fim, o volume 4 se intitula "E quando a política está em andamento? Avaliação *ex post*!". Ele trata de todo o processo de avaliação de uma política pública em andamento ou já concluída, composto de quatro etapas: avaliação de desenho, avaliação de processos, avaliação de impacto e avaliação de custo-benefício. Essas etapas são abordadas em capítulos separados, nos quais são apresentadas as metodologias específicas de cada tipo de avaliação. Espera-se que ao final de uma avaliação *ex post*

seja possível responder às seguintes questões: (i) o desenho da política é consistente com seus objetivos e adequado à solução dos problemas-alvo? (ii) as atividades executadas são consistentes com o desenho? (iii) o programa tem impactos causais sobre as dimensões esperadas? Quais as magnitudes desses impactos? Ademais, o programa tem impactos causais não esperados, sobre outras dimensões? e (iv) qual o custo necessário para se alcançar o resultado obtido? Os benefícios gerados pelos impactos causais da política compensam os custos incorridos?

Espera-se que este Guia para Avaliar Políticas Públicas colabore para a prática de tomadas de decisões baseadas em evidências. O seu objetivo é contribuir para a disseminação da cultura de monitoramento e avaliação das políticas públicas. Embora elaborado para a utilização imediata dos usuários do SiMAPP, ele vai mais além. É um guia útil para gestores, especialistas ou mesmo qualquer pessoa que esteja preocupada com o uso eficiente dos recursos públicos para a melhoria do bem-estar da população.

André Portela Souza
Diretor do
FGV EESP Clear

Como utilizar este Guia?

Inicialmente, são contextualizados o tema da avaliação de políticas públicas e os benefícios que ela traz, bem como as questões-chave das avaliações *ex post*. Essa primeira leitura auxilia a compreensão da importância da avaliação no ciclo das políticas públicas. A partir daí, são apresentadas as etapas a serem percorridas. São elas:

1 Avaliação de desenho

É analisada a consistência do desenho da política, o que requer elaboração de um Modelo Lógico, caso não tenha sido feito à época da sua formulação. A partir disso, devem ser examinadas a adequação das ações ao problema social a ser combatido e a plausibilidade lógica das hipóteses sobre como o programa atingiria os resultados previstos.

2 Avaliação de processos

Análise da coerência entre o que foi pretendido pelo desenho da política e como ela foi operacionalizada na prática. A implementação e o funcionamento da política são examinados, com foco nos recursos utilizados, atividades realizadas e produtos entregues.

3 Avaliação de impacto

São apresentadas metodologias para identificar e mensurar a relação de causalidade entre a política implementada e as mudanças observadas nos indicadores de resultado e impacto. Para isso, são feitas comparações entre: (i) o que se observa na realidade, uma vez implementada a política; e (ii) o que se espera que se observaria caso a política não fosse implementada (contrafactual). Os métodos utilizados podem ser experimentais ou não experimentais.

4 Avaliação de custo-benefício e custo-efetividade

Os impactos gerados pela política devem ser avaliados em relação ao seu custo. As avaliações de custo-benefício confrontam impactos mensuráveis em termos monetários com os custos da política. Já as avaliações custo-efetividade, confrontam impactos não monetizáveis com os custos da política.

1 Introdução

Políticas públicas são as ações do Estado para resolver ou tratar problemas da sociedade. O entendimento do que é um problema passa pela comparação da situação vivida com a situação ideal desejada. Assim, essas ações, que podem tomar forma de programas, leis, incentivos econômicos, entre outros, têm como objetivo prover a melhoria de algum cenário e efetivar direitos.

A atuação dos governos é regulada por um conjunto de princípios e normas jurídicas. Dessa forma, muita atenção deve ser dada à conformidade legal dos processos aplicados à execução das políticas públicas e demais atividades. No entanto, há outros aspectos fundamentais a serem observados. A efetividade das ações estatais em mitigar ou resolver os problemas que as motivaram é um elemento pouco analisado no Brasil, mas que diz respeito à própria razão de ser das políticas. Dessa forma, a análise dos resultados das intervenções na vida dos cidadãos é primordial.

> “As práticas de monitoramento e de avaliação (M&A) produzem evidências sobre a efetividade ou não das políticas públicas”

As práticas de monitoramento e de avaliação (M&A) produzem evidências sobre a efetividade ou não das políticas públicas. Assim, podem ser utilizadas como instrumentos poderosos para a tomada de decisão dos gestores. Com o uso dessas técnicas, o governo pode conhecer os resultados das suas intervenções, levando à aprendizagem institucional e a escolhas mais assertivas sobre em quais políticas investir mais recursos e sobre quais devem passar por ajustes. Se pudéssemos resumir todo propósito das práticas de monitoramento e de avaliação em uma frase, ela seria “prover informações úteis e bem fundamentadas” ou “fornecer evidências para a tomada de decisão”. Dessa forma, as metodologias de M&A têm o potencial de elevar o padrão de qualidade do serviço e a eficiência do gasto público. Além disso, têm um papel muito importante em fortalecer a prestação de contas das ações do governo (*accountability*) e aumentar a transparência.

Não só no Brasil, mas em todo o mundo, crescem as pressões por maior responsabilização dos governos e das organizações em transmitir confiança e conquistar

reconhecimento pelo desempenho e pela efetividade de suas ações. As práticas de M&A fazem parte, por exemplo, das diretrizes de atuação da Organização das Nações Unidas (ONU), de países como Chile, Reino Unido, Coreia do Sul, Estados Unidos, México e Colômbia e dos modelos de gestão para resultados. O monitoramento do aquecimento global, da produtividade dos trabalhadores de uma indústria, dos casos de doenças contagiosas em áreas de risco, bem como o parecer quanto à criação, continuidade ou aperfeiçoamento de uma política de incentivo à educação infantil são apenas alguns exemplos de aplicações práticas desse mecanismo na análise de questões complexas, e que fazem toda a diferença para a qualidade das decisões dos gestores.

Nesse cenário, o Governo do Estado do Espírito Santo está dando um passo pioneiro no Brasil, ao criar o Sistema de Monitoramento e Avaliação de Políticas Públicas (SiMAPP), que trará as práticas de M&A para dentro do ciclo de planejamento do governo, de maneira sistemática e padronizada[1]. Esse Sistema foi implementado por meio da Lei Estadual nº 10.744, de 05 de outubro de 2017 (ESPÍRITO SANTO, 2017) e seus resultados irão possibilitar um debate mais qualificado sobre a priorização de políticas públicas, subsidiando o Estado na elaboração e na revisão do planejamento. O desafio do governo para esta e para as próximas gerações não é apenas conter a expansão do gasto público, mas avaliar onde ele é mais produtivo, buscando fazer mais com menos recursos e priorizando a eficiência das políticas públicas.

1.1 Avaliação de políticas públicas

Avaliações são conduzidas por uma variedade de razões práticas: preocupações sobre as necessidades do público-alvo e se ele está sendo adequadamente atendido, preocupações sobre a gestão e operação do programa, a eficácia dos serviços, se o programa está tendo o impacto desejado e se seus benefícios

[1] Para mais informações sobre o SiMAPP, ver IJSN (2018), disponível em: <http://www.ijsn.es.gov.br/component/attachments/download/6376>.

compensam seus custos, entre diversas outras. Fundamentalmente, o propósito de qualquer avaliação consiste em fornecer informações aos gestores de políticas públicas e habilitá-los a tomar as melhores decisões diante da concepção, continuidade, reformulação ou extinção de políticas públicas.

Dessa forma, a elaboração de uma avaliação deve surgir da definição das questões que buscará responder. Várias questões de interesse podem aparecer no decorrer do ciclo de uma política pública, a depender de seu estágio de maturidade (antes, durante ou depois de sua implementação). Entretanto, uma avaliação deve ser objetiva quanto ao conjunto de informações que pretende levantar, uma vez que a natureza das questões da avaliação influencia diretamente o tipo de avaliação a ser conduzida e a metodologia a ser aplicada.

Boxe 1

Exemplos de questões que uma avaliação pode ajudar a responder:

- Qual a natureza do problema que se quer enfrentar? A atuação do Estado é justificada?
- A intervenção está sendo bem implementada? As ações pretendidas estão sendo realizadas?
- Quem está sendo beneficiado pela intervenção?
- Qual foi a performance da intervenção? Os impactos esperados foram atingidos?
- Qual o custo da intervenção? O custo da política é razoável em relação a seus benefícios?

Uma forma de classificar diferentes tipos de avaliação é considerar o momento em que a avaliação é conduzida. Uma avaliação *ex ante* é uma avaliação que ocorre antes da implementação da política, partindo-se do diagnóstico do problema social a ser enfrentado até o momento em que o planejamento e as escolhas são

transformados em prática. Uma avaliação *ex post* é, por sua vez, realizada após a implementação da política. Diversos tipos de avaliação *ex post* podem ser desenvolvidos, com finalidades diversas e complementares, a depender das questões para as quais se buscam respostas. A análise executiva, por exemplo, é uma avaliação que gera uma visão global do desempenho da política pública já implementada, mesclando várias metodologias e verificando, inclusive, a necessidade de elaboração de avaliações *ex post* mais aprofundadas. Por isso, cada um dos volumes deste Guia se dedica a um tipo de avaliação.

1.2. Avaliação *ex post*

Avaliações *ex post*[2] são um conjunto de metodologias de avaliação aplicadas quando a política já está em andamento (ou foi até finalizada). A Lei estadual nº 10.744, publicada no Diário Oficial do Espírito Santo em 05 de outubro de 2017 (ESPÍRITO SANTO, 2017), que cria o SiMAPP, define a linha de avaliação de políticas públicas em andamento como aquela em que se "avalia o desenho (objetivos, componentes de produção, população alvo, beneficiários efetivos, período de execução, âmbito territorial, fontes de financiamento e outros aspectos importantes que caracterizam o programa), a gestão e os resultados do programa, analisando a consistência do desenho e dos resultados esperados".

Assim, uma gama muito ampla de perguntas de avaliação pode ser formulada a respeito do desenho, dos processos ou ainda dos resultados e custos de uma política. Em todos os casos, quando são feitas as perguntas certas, a obtenção de respostas no processo de avaliação *ex post* identifica as principais falhas que afetam o desempenho da política ou os principais fatores de sucesso. Por exemplo, quando é identificado um problema entre os insumos e atividades ou entre as atividades e produtos, diz-se que houve um erro ou falha de implementação. Quando, por outro lado, os insumos aplicados, as atividades executadas e os produtos

[2] A expressão ***ex post*** vem do Latim e é utilizada para designar uma análise que é feita após a ocorrência do fato.

gerados correspondem ao planejado e ainda assim os resultados e impactos esperados não são obtidos, identifica-se que a falha possivelmente está localizada na Teoria do Programa.

Conhecer o tipo de erro que obstrui a obtenção de resultados por uma política é crucial para seu futuro. Se o que falhou foi o processo de implementação, ações para melhoria de seu desempenho deverão focar na revisão dos recursos empregados e atividades realizadas, corrigindo os erros com alterações de práticas e estratégias de gestão de equipe e de processos de implementação. No entanto, se a falha for teórica, poderá ser necessário um redesenho parcial ou completo do programa em questão, o que frequentemente gera implicações políticas. Falhas e inconsistências no desenho de uma política são identificadas a partir de uma **avaliação de desenho**, apresentada no capítulo 2.

“Conhecer o tipo de erro que obstrui a obtenção de resultados por uma política é crucial para seu futuro”

Perguntas de **avaliação sobre o processo** de implementação envolvem, por exemplo, compreender se a quantidade de insumos previstos foi suficiente para a execução das atividades da política, se os parceiros executores e fornecedores cumpriram o que havia sido acordado, se o número de produtos entregues seguiu o planejamento inicial, entre outras. Desta forma, estão relacionadas aos componentes de *insumos, atividades* e *produtos* do Modelo Lógico, e a obtenção de suas respostas remete, então, a uma avaliação de processos, detalhada no capítulo 3.

Perguntas de **causa e efeito** são aquelas voltadas às mudanças geradas pelo programa. Há diversos fatores externos em ação durante a execução de uma intervenção, de modo que não se pode atribuir automaticamente ao programa a responsabilidade por essas mudanças (positivas ou negativas). Pode-se perguntar, por exemplo, se uma política causou exclusivamente mudanças (diminuições ou aumentos) nas taxas de mortalidade infantil, na evasão escolar ou no número de crimes per capita em uma determinada região, e qual o tamanho desses efeitos. Assim, este tipo de pergunta

remete à necessidade de uma **avaliação de impacto**, por meio da qual se pode identificar com precisão se houve uma relação de causalidade entre a intervenção executada e as mudanças observadas, e qual a magnitude dessa relação. Esse tema é explorado no capítulo 4.

Por fim, os custos incorridos pela política são explorados e analisados a partir de **avaliação de custo-benefício e custo-efetividade**, detalhadas no capítulo 5. Esses tipos de avaliação permitem responder a perguntas relacionadas ao nível de investimento necessário para se alcançar os resultados obtidos, além do quanto os investimentos realizados pela política são compensados pelos benefícios obtidos através dela.

2

Avaliação de desenho

De forma geral, pode-se dizer que, quando uma política pública falha em alcançar seus objetivos, isso se dá por problemas em seu desenho ou em seu processo de implementação. A importância de se iniciar uma avaliação *ex post* com uma análise da teoria da política em foco surge para que possíveis problemas no seu desenho sejam minimizados. Quando o desenho apresenta problemas consideráveis, já detectáveis por uma análise de seu Modelo Lógico, a probabilidade de alcance dos resultados esperados é baixa. Ainda que uma implementação conduzida de forma problemática possa levar ao fracasso da política em combater o problema a que ela se propôs tratar, uma intervenção com problemas no desenho terá grandes chances de não alcançar seus objetivos mesmo quando perfeitamente implementada.

Conforme apresentado no Capítulo 3 do volume "A Política é Nova? Avaliação *ex ante*!" deste Guia, o Modelo Lógico é uma forma visual e sucinta de se expor a chamada Teoria do Programa, definida pela sequência lógica das relações de causa-efeito entre cada etapa planejada da política. Através dele, será apresentada a lógica causal entre os componentes de um programa, projeto ou política, via a descrição de cinco etapas - insumos, atividades, produtos, resultados e impactos - e da sequência lógica que as une, explicitando os mecanismos por meio dos quais se visa a obter os resultados esperados no curto, médio e longo prazo.

Uma avaliação de desenho tem por objetivo, então, analisar a consistência lógica da Teoria do Programa e sua adequação às necessidades ou problemas sociais que a política busca combater (J-PAL, 2016). Para isso, são necessárias duas etapas, apresentadas nas seções a seguir com base nas discussões de Rossi et al. (2003, cap. 5): (i) a elaboração do Modelo Lógico; e (ii) a análise da teoria do programa explicitada pelo Modelo Lógico construído.

2.1. Elaboração do Modelo Lógico

Idealmente, a elaboração do Modelo Lógico de uma política deve acontecer paralelamente à definição de seu desenho, ainda antes de sua implementação, durante uma avaliação *ex ante*. Dessa forma, torna-se mais fácil garantir a consistência da política, orientando todas as etapas de implementação de forma estruturada, visando sempre aos resultados e impactos pretendidos. Esse processo também permite que se indique os dados a serem monitorados durante e após a intervenção, para que se possa verificar seu desempenho.

Entretanto, por vezes é comum que um programa seja implementado, ou mesmo inteiramente executado, sem a elaboração prévia de um Modelo Lógico. Nesses casos, essa ferramenta ainda é importante, já que somente a partir da clareza quanto à Teoria do Programa e aos objetivos de curto, médio e longo prazo será possível embasar a análise do desenho da política. Além disso, a partir desse instrumento é possível encontrar as perguntas de avaliação *ex post* específicas para a política em questão e, dessa forma, elucidar os tipos de avaliação e metodologias a serem empregados para a obtenção de tais respostas.

A Teoria do Programa deve ser identificada da forma como foi pretendida à época da elaboração. Assim, o Modelo Lógico deve refletir as expectativas dos formuladores da política acerca das atividades planejadas e dos resultados que ela pretendia obter.[3] Vale notar que, mesmo nos casos em que um Modelo Lógico já exista, é importante verificar o seu nível de detalhamento e, então, realizar os ajustes necessários. Um Modelo Lógico pouco informativo, que não reflita a Teoria do Programa, dificulta a avaliação de desenho da política. Assim, as informações necessárias à elaboração do Modelo Lógico podem ser divididas em três blocos (ROSSI *et al.*, 2003):

(i) Resultados esperados: dizem respeito às mudanças que a política pretende causar sobre sua população-alvo. São relacionados, portanto, aos componentes de *resultados* e *impactos* do Modelo Lógico;

[3] Em muitos casos, é natural que tenham acontecido adaptações, ao longo do tempo, do que havia sido planejado em relação ao que foi de fato implementado. Esse descasamento entre o Modelo Lógico pretendido e a operação da política é analisado a partir de uma avaliação de processos, explorada no Capítulo 3 deste volume.

(ii) Operações: são relacionadas ao planejamento de como a política deve ser implementada e operacionalizada, considerando os componentes de *insumos, atividades* e *produtos* do Modelo Lógico;

(iii) Lógica da política: diz respeito às conexões causais entre os componentes do Modelo Lógico e as hipóteses relacionadas, obtidos nos blocos acima. Essas informações são particularmente importantes para que, uma vez descrita a Teoria do Programa, a consistência do desenho da política possa ser analisada.

De forma a refletir a Teoria do Programa concebida pelos formuladores, essas informações podem ser obtidas a partir de diversas fontes, como registros e documentos relacionados à política e entrevistas realizadas com gestores ou demais servidores que atuam na implementação da política. Os Boxes A, B e C apresentam recomendações gerais sobre coleta de dados a partir da revisão de documentos, realização de entrevistas e de grupos focais, respectivamente. A Tabela 1 traz o *template* para a elaboração de um Modelo Lógico.

Template disponível em:
www.ijsn.es.gov.br/CMA/GUIA

Escaneie no seu celular:

Tabela 1 | Modelo Lógico

Insumos	Atividades	Produtos	Resultados	Impactos

Note que, até agora, o esforço realizado foi no sentido de descrever objetivamente o desenho da política. Entretanto, é possível que já nessa etapa sejam identificadas dificuldades. Por exemplo, é possível que os objetivos da política não estejam claramente definidos, ou que haja alguma discordância por parte de atores-chave sobre a forma como as atividades plane-

jadas pela política devem gerar os resultados esperados e, mais ainda, quais devem ser realmente esses objetivos. Identificar esses problemas também é parte do escopo da avaliação de desenho. De qualquer forma, uma vez em posse de um Modelo Lógico que reflita a Teoria do Programa, pode-se partir para sua análise. Nesse sentido, deve-se avaliar sua razoabilidade enquanto plano para o funcionamento da política. A seção a seguir aborda esse assunto.

Boxe A

Coleta de dados: Revisão de documentos

A revisão de documentos é uma maneira de coletar dados a partir da revisão de informações já existentes. Os documentos podem ser relatórios, registros do programa, avaliações de desempenho, propostas de financiamento, atas de reuniões, newsletters, materiais de marketing, entre outros.

Para que usar a revisão de documentos?

Utiliza-se a revisão de documentos para reunir informações básicas sobre a política. Essa ferramenta auxilia a entender a história, a filosofia e a operação do programa que se pretende avaliar, assim como ajuda a formular perguntas para entrevistas, questionários e grupos focais. A revisão de documentos é útil para determinar se a implementação do programa reflete o que foi planejado inicialmente.

Como planejar e conduzir a revisão de documentos?

É importante limitar a revisão apenas aos documentos que respondem às perguntas de interesse da avaliação, além de entender como e por que os documentos foram produzidos. Durante a condução da revisão, pode ser útil criar um formulário de coleta de dados para resumir os dados coletados durante as análises de documentos.

Quais são as vantagens da revisão de documentos?

Revisão de documentos é um método relativamente barato e que pode constituir uma boa fonte de informações de background. Fornece uma visão dos bastidores de um programa que pode não ser observada diretamente.

Quais são as desvantagens da revisão de documentos?

As informações podem estar desorganizadas, indisponíveis, desatualizadas, incompletas ou imprecisas. Trata-se de um método que é suscetível ao viés de sobrevivência (documentos podem ter sido perdidos ou substituídos ao longo do tempo, por exemplo). Pode ser um processo demorado de coleta, revisão e análise.

Fonte: adaptado de *US Centers for Disease Control and Prevention*, "Data Collection Methods for Evaluation: Document Review" Evaluation ETA Evaluation Briefs, No. 18, Janeiro de 2009. Disponível em: <www.cdc.gov/healthyyouth/evaluation/pdf/brief18.pdf>

Boxe B

Coleta de dados: Entrevistas

A entrevista é um método de coleta de dados quantitativos ou qualitativos através de perguntas feitas oralmente. Em geral, perguntas quantitativas são fechadas e objetivas, com opções de respostas específicas que podem ser categorizadas, enquanto perguntas qualitativas são abertas, ou seja, podem ser respondidas com as próprias palavras do entrevistado.

Para que usar entrevistas?

Utiliza-se entrevistas para obter informações mais detalhadas sobre a percepção, *insights*, atitudes, experiência ou crenças. A entrevista é útil para reunir perspectivas subjetivas dos entrevistados e para avaliar diferenças individuais entre suas experiências e resultados. Pode também ser utilizada como um *follow-up* para outros métodos (por exemplo, para auxiliar na interpretação de dados quantitativos coletados através de questionários).

Como planejar uma entrevista?

- Determinar seu escopo;
- Definir as questões de avaliação a serem respondidas;
- Desenvolver um guia/questionário de entrevista;

- Selecionar a quantidade e o tipo de pessoas a serem entrevistadas;
- Treinar os entrevistadores;
- Garantir a confidencialidade dos entrevistados e informá-los sobre como isso será feito;
- Realizar um pré-teste (piloto) do guia/questionário de entrevista.

Como conduzir uma entrevista?

É importante que o entrevistado se sinta à vontade para responder às perguntas com honestidade, motivo pelo qual a construção de confiança e empatia são essenciais. O entrevistador deve se manter com comportamento neutro, não demonstrando reações às respostas obtidas, além de manter-se sempre no controle da entrevista para garantir que ela não fuja do escopo definido.

Quais são as vantagens da entrevista?

Entrevistas são úteis para obter *insights*, contexto sobre um determinado tópico, coletar citações e histórias. Permitem que o entrevistado descreva o que é importante sob sua visão.

Quais são as desvantagens da entrevista?

Entrevistas são suscetíveis ao viés do entrevistador, em que o responsável por conduzir a entrevista pode influenciar as respostas obtidas, mesmo não intencionalmente. É um método relativamente demorado e caro comparado a outros métodos de coleta de dados. A depender do escopo da entrevista, da forma como for planejada e conduzida, o entrevistado pode se sentir desconfortável (por exemplo, para responder a questões de cunho pessoal).

Fonte: adaptado de ***US Centers for Disease Control and Prevention***, "Data Collection Methods for Evaluation: Interviews" Evaluation ETA Evaluation Briefs, No. 17, Janeiro de 2009. Disponível em: <www.cdc.gov/healthyyouth/evaluation/pdf/brief17.pdf>

Boxe C

Coleta de dados: Grupo focal

Grupo focal é um método de coleta de dados qualitativos através de uma entrevista com um grupo de seis a doze pessoas, que podem compartilhar de características semelhantes e interesses comuns ou de características distintas. Nele, o facilitador promove um ambiente que incentive os participantes a compartilharem suas percepções e pontos de vista.

Para que usar grupos focais?

Os grupos focais são úteis para reunir perspectivas subjetivas dos entrevistados e para obter informações mais detalhadas sobre a percepção, *insights*, atitudes, experiências ou crenças dos indivíduos. Podem ser úteis para fomentar interpretações acerca de dados quantitativos.

Como planejar um grupo focal?

Os dois principais componentes do planejamento de entrevistas com grupos focais envolvem o desenvolvimento de um guia de entrevista, que servirá como um roteiro, e a decisão sobre o número e o perfil dos participantes.

Como conduzir uma entrevista de grupo focal?

É preciso de um facilitador para guiar o grupo através da discussão; alguém para tomar notas que não interaja com o grupo, apenas observe e; um técnico para gravar a entrevista com o grupo focal, quando cabível. Os grupos focais podem ser realizados pessoalmente ou por conferência *online*.

Quais são as vantagens dos grupos focais?

Grupo focal é um método rápido e relativamente fácil de construir. A dinâmica em grupo pode fornecer informações úteis que a entrevista individual não fornece, bem como *insights* sobre um tópico que podem ser mais difíceis de obter por meio de outros métodos.

Quais são as desvantagens dos grupos focais?

Grupos focais são suscetíveis ao viés do facilitador, em que o responsável por conduzir o grupo focal pode influenciar as respostas obtidas, mesmo não intencionalmente. A discussão pode ser dominada ou desviada

por alguns indivíduos. A análise de dados pode ser complexa por demandar julgamentos subjetivos. Além disso, não fornece informações válidas a nível individual e a informação não é, em geral, representativa para outros grupos.

Fonte: adaptado de *US Centers for Disease Control and Prevention*, "Data Collection Methods for Evaluation: Focus Groups" Evaluation ETA Evaluation Briefs, No. 13, Julho de 2008. Disponível em: <www.cdc.gov/healthyyouth/evaluation/pdf/brief13.pdf>.

Exemplo 1

Curso de capacitação profissional customizado

Com o objetivo de promover a produtividade média por meio do aumento da empregabilidade e da renda, o governo de um determinado estado desenvolveu uma política pública em que oferta um curso de capacitação profissional customizado[4]. O curso tem duração de 1 ano com carga horária de 800 horas e é articulado ao Ensino Médio regular, sendo oferecido na modalidade concomitante (realizado no contraturno) e tendo como público-alvo, portanto, jovens de 15 a 19 anos. O curso é oferecido apenas em escolas estaduais localizadas em municípios com até 100.000 habitantes. Os alunos podem escolher a área na qual irão cursar a capacitação profissional (ex: administração, tecnologia da informação, contabilidade).

O currículo do curso é composto pelos seguintes módulos:

1. Capacitação profissional na área escolhida Parte I;
2. Capacitação profissional na área escolhida Parte II;
3. Educação digital;
4. Habilidades socioemocionais;
5. Realização de estágio na área escolhida.

A customização do curso está na inclusão no currículo: (i) do módulo de educação digital, que busca promover o acesso à informática; e (ii) do módulo de habilidades socioemocionais, que trata do desenvolvimento desse tipo de habilidades com foco no mercado de trabalho e apresenta três eixos principais: comportamento, cidadania e empregabilidade. As principais habilidades trabalhadas são: como se portar no mercado de trabalho, como trabalhar em equipe, proatividade e resolução de problemas.

[4] Todo o conteúdo do presente Boxe é fictício e foi elaborado apenas para exemplificar essa seção do Guia.

O curso encontra-se atualmente em sua segunda edição e o governo está considerando propor algumas reformulações no desenho da política. Para embasar esse processo de tomada de decisão, decidiu-se conduzir uma avaliação de desenho *ex post* da política em questão.

A primeira etapa dessa avaliação é a elaboração do Modelo Lógico da política em análise. Como este não foi elaborado no momento de concepção da política, antes de sua implementação, foi preciso elaborá-lo no próprio contexto da avaliação de desenho *ex post*. Para isso, foram utilizadas as seguintes formas de coleta de dados e de fontes de informações:

- Revisão de Documentos: foram realizados levantamentos da proposta pedagógica do programa, de currículos planejados, de relatórios desenvolvidos para informar à Secretaria Estadual de Educação sobre a execução do mesmo, de relatórios dos recursos (financeiros e humanos) empregados no programa, de dados sobre caracterização das escolas ofertando o programa (perfil socioeconômico dos alunos, equipamentos e infraestrutura das escolas, número de funcionários e perfil dos docentes), de fichas de cadastros dos alunos inscritos, de registros das avaliações dos alunos, de atas de reuniões e de materiais de divulgação.
- Entrevistas: foram realizadas diversas entrevistas com atores-chave envolvidos no programa, tendo sido utilizados roteiros específicos para cada tipo de ator entrevistado (ex: gestores, professores e alunos). As informações coletadas nas entrevistas dividem-se em quatro temas: (i) Alunos; (ii) Professores; (iii) Gestão; e (iv) Repasses Financeiros. O objetivo desse processo foi compreender todos os detalhes processuais, logísticos e operacionais do programa, incluindo o processo de recrutamento dos alunos, sua matrícula e acompanhamento de rendimento escolar, contratação de professores, elaboração de currículos e oferecimento do curso.
- Grupos focais: foi realizado um grupo focal, que foi conduzido por um facilitador experiente, que também colaborou na elaboração do guia de entrevista utilizado. O grupo contou com 8 participantes, sendo eles representantes de servidores envolvidos no planejamento, gestão e execução do programa, diretores, coordenadores pedagógicos e professores das escolas. Foram discutidas as percepções dos participantes sobre o curso, seus objetivos e a capacidade de alcançá-los.

Com base nos dados e informações coletados nas atividades descritas acima, foi desenvolvido o seguinte Modelo Lógico do programa.

Tabela 2 | Modelo lógico do curso de avaliação profissional customizado

Insumos	Atividades	Produtos	Resultados	Impactos
Recursos Físicos (escolas, salas, materiais etc.) Recursos Humanos (professores, secretárias, pessoal adminis-trativo etc.) Recursos Financeiros (Estaduais) Participantes (jovens de 15 a 19 anos, com ensino fundamental completo)	Elaborar currículos articulando conteúdos gerais e técnicos Oferecer aulas de capacitação	Currículos articulados elaborados Aulas de capacitação profissional ofertadas	Conhecimentos gerais e técnicos adquiridos pelos alunos Habilidades técnicas e socioemocionais desenvolvidas pelos alunos Conclusão do curso pelos alunos	Aumento da empregabilidade dos alunos Aumento da renda do trabalho dos alunos Aumento da produtividade média no Estado

2.2. Análise da Teoria do Programa

A pertinência da teoria de um programa deve ser estudada com cuidado, por se tratar de um processo que usualmente envolve um componente significativo de senso comum. Ainda assim, é sempre necessário que sejam examinadas todas as hipóteses das quais depende o desempenho de uma política, incluindo considerações quanto ao contexto em que ela é conduzida. Mesmo uma intervenção que pareça simples diante de uma análise inicial pode passar por considerações mais complexas a depender, por exemplo, da região onde é implementada. Em outras palavras, é possível que alguns pressupostos sejam válidos para determinado contexto, por exemplo, mas não para outros, e essas particularidades devem ser levadas em consideração pela Teoria do Programa.

Referências de experiências comparáveis à política em questão, em outros estados ou até mesmo países, que tenham gerado evidências acerca do que funciona e do que não funciona para aquele tipo de intervenção, constituem um importante passo para que a consistência da Teoria do Programa seja analisada. Em particular, para políticas que tenham passado por avaliação *ex ante*, é possível que tais referências já tenham sido identificadas e contribuído para seu desenho. A comparação da racionalidade de um programa a experiências de sucesso pode auxiliar na demonstração da credibilidade de determinadas hipóteses, além de contribuir para a identificação dos elementos com maior potencial para tornarem-se os pontos de risco de uma intervenção.

Adequação às necessidades sociais

Em relação à adequação da política para contribuir com as necessidades de sua população-alvo, é importante considerar o diagnóstico do problema social que tenha originado o desenho da política. Quanto maior a compreensão das especificidades do problema, incluindo a identificação de suas causas potenciais, de suas particularidades regionais e das diferentes populações afetadas por ele, maior a chance de que a lógica da política tenha levado em consideração todos os fatores que podem afetar seu desempenho e, portanto, mais robusta pode-se considerar a Teoria do Programa.

Dessa forma, o papel da avaliação de desenho é o de verificar em que medida a política é relevante para combater o problema social identificado. Em outras palavras, é necessário contrastar as mudanças pretendidas pela política (seus resultados esperados) e as necessidades sociais em questão. Nesse sentido, devem ser consideradas questões como a especificidade e clareza dos objetivos de curto, médio e longo prazo e a facilidade de acesso, por parte da população-alvo, aos bens e serviços oferecidos pela política.

Plausibilidade lógica

Por fim, também deve-se analisar a plausibilidade das conexões lógicas da Teoria do Programa. Para isso, além da razoabilidade lógica de senso comum, muitas vezes pode ser útil consultar especialistas sobre o tema em questão, para verificar a razoabilidade das hipóteses causais da política em termos de evidências científicas da área.

No âmbito da lógica da política, apesar de não constituírem uma lista exaustiva, há certos aspectos que devem ser considerados:

1) O nível de definição, clareza, e concretude dos resultados e impactos esperados documentados no Modelo Lógico, que devem ser suficientes para que se possa averiguar sua obtenção com facilidade;

2) O nível de realismo dos resultados e impactos esperados do programa. Expectativas demasiadamente altas em relação à performance de um programa ou a sua capacidade de impactar esferas fora de seu controle direto não apontam para hipóteses robustas, por exemplo;

3) O nível de razoabilidade e plausibilidade das relações causais entre os componentes da política. Onde forem identificadas lacunas, é importante considerar se estas poderiam ser preenchidas a partir de um maior detalhamento do Modelo Lógico;

4) A adequação dos procedimentos relacionados ao ciclo dos serviços propostos, passando pela identificação e seleção de beneficiários entre o público-alvo, implementação do serviço e sua execução por tempo suficiente para a entrega dos produtos previstos.

Ao final da avaliação de desenho, espera-se que eventuais falhas da Teoria do Programa tenham sido identificadas. A partir das deficiências encontradas, é importante que sejam propostas possíveis soluções. Por exemplo, problemas relacionados à falta de clareza dos objetivos da política por vezes podem ser solucionados a

partir de um maior detalhamento do Modelo Lógico, desde que haja concordância entre os responsáveis pela política quanto a quais devem ser esses objetivos. Por outro lado, se forem verificadas falhas mais complexas, como problemas graves de plausibilidade lógica ou de adequação da política às necessidades sociais envolvidas, um redesenho mais aprofundado pode ser necessário.

Vale notar que a concordância por parte dos atores-chave da política – sejam seus gestores ou demais níveis de governo – em relação a quaisquer mudanças propostas, tanto acerca de seus objetivos quanto de seu desenho, envolve importantes considerações políticas, que devem ser devidamente acordadas entre todos os envolvidos.

Exemplo 1

Curso de capacitação profissional customizado

Uma vez elaborado o Modelo Lógico da política pública que oferta um curso de capacitação profissional customizado, conforme discutido e apresentado previamente no Boxe de exemplo da seção 2.1 de avaliação de desenho *ex post*, passou-se para a etapa de análise da teoria do programa.

A partir de todas as informações coletadas e do Modelo Lógico elaborado, em relação à adequação às necessidades sociais, foi constatado que:

Quanto à plausibilidade lógica, foi constatado que:

- A partir das informações levantadas na proposta pedagógica do programa, verificou-se que: (i) havia sido planejada a elaboração de currículos articulando os conteúdos gerais (ensino regular) e técnicos; mas (ii) não havia sido planejada a integração entre os currículos propostos para os módulos de "capacitação profissional na área escolhida" e "habilidades socioemocionais", de forma que esses conteúdos fossem trabalhados de maneira transversal. Esse aspecto foi considerado como uma falha lógica no

desenho da política, uma vez que há evidências na literatura de que essa integração é muito importante para o pleno desenvolvimento de habilidades socioemocionais pretendidas.

Recomendação associada apresentada na avaliação de desenho: consultar especialistas sobre quais atividades específicas podem ser desenvolvidas para promover a integração do currículo dos módulos de "capacitação profissional na área escolhida" e "habilidades socioemocionais".

- Segundo relatos dos gestores do programa, outro fator que consideram ser relevante para explicar as altas taxas de evasão observadas no curso é o módulo de realização de estágio na área escolhida, que deve ser cumprido para que o aluno obtenha o certificado de conclusão de curso. Na análise dos gestores, como trata-se de um curso de capacitação profissional articulado ao Ensino Médio regular na modalidade concomitante, os alunos têm dificuldade de encontrar vagas de estágio para as quais podem se candidatar, pois eles geralmente já têm as manhãs e tardes ocupadas com as aulas do ensino regular e do curso de capacitação. Por não encontrarem vagas disponíveis, os alunos têm dificuldade para completar esse módulo e acabam desistindo do curso. Esse problema está relacionado ao fato de que, na grade do curso, não há um semestre reservado especificamente para esse módulo, que deve ser realizado pelo aluno simultaneamente a outros módulos do curso.

  Recomendação associada apresentada na avaliação de desenho: (i) reformular a grade do curso, de forma a deixar um trimestre reservado exclusivamente para a realização do módulo de estágio na área escolhida; (ii) analisar a possibilidade de estabelecimento de parcerias com possíveis empregadores para o oferecimento de vagas de estágio exclusivas para alunos do curso; (iii) analisar a possibilidade de flexibilização da exigência de cumprimento do módulo de estágio na área escolhida para a emissão de certificados de conclusão do curso.

- Com base nas informações obtidas durante a entrevista de grupo focal, foi identificado que um dos impactos esperados pelo programa diz respeito ao "aumento da produtividade média no Estado", conforme apresentado no Modelo Lógico elaborado. O problema constatado nesse caso é relacionado ao nível de realismo desse objetivo. É necessário consi-

derar a escala do programa, que atende à subpopulação de alunos de escolas estaduais em municípios relativamente pequenos (de até 100.000 habitantes), frente à magnitude do impacto esperado.

Recomendação associada apresentada na avaliação de desenho: revisar as expectativas sobre o impacto esperado tendo em vista o alcance do programa, limitando-o à sua população-alvo.

3 Avaliação de processos

A análise das operações de um programa ou política constitui o que é chamado de avaliação de processos (também conhecida como avaliação de implementação). Nesse tipo de avaliação, o objetivo é analisar a implementação e o funcionamento de uma política para identificar os fatores que promovem ou obstruem sua efetividade, a partir de considerações acerca dos recursos utilizados, das atividades realizadas e dos produtos entregues. Em meio à complexidade das operações em que se baseia uma política pública, é importante saber por que um resultado é alcançado ou não, que grupos conseguem obter esses resultados e sob que circunstâncias ele ocorre.

Os usos que se pode fazer de uma avaliação de processos são vários. De modo geral, conforme discutido por Saunders *et al*. (2005), ela pode servir para o aperfeiçoamento de uma política (avaliação formativa) ou para a tomada de decisão sobre adoção ou expansão de uma política (avaliação somativa). Além disso, a análise da implementação da política pode mitigar a chance de se cometer o chamado "erro do tipo III": avaliar um programa que não foi corretamente implementado[5] (BASCH *et al*., 1985).

Nesse sentido, seguindo a discussão de Rossi *et al*. (2003), as perguntas respondidas por esse tipo de avaliação são usualmente de dois tipos: (i) utilização de serviços, que consiste em verificar se o público-alvo[6] pretendido recebe os bens e serviços planejados pela política; e (ii) organização da política, que consiste em comparar o planejamento inicial com o que de fato é implementado na prática.

A escolha das perguntas a serem respondidas pela avaliação de processos, assim como das metodologias empregadas para respondê-las, dependerá das especificidades de cada política. Portanto, é essencial que se tenha conhecimento adequado da teoria da política avaliada, de forma a facilitar a identificação das diferenças entre o funcionamento prático da política e o seu planejamento inicial. Além disso, conforme discutido por Saunders *et al*. (2005), é importante também

[5] O nome faz analogia aos erros de "tipo I" e "tipo II" da estatística: rejeitar uma hipótese quando ela é verdadeira e não rejeitar uma hipótese quando ela é falsa, respectivamente.

[6] O público-alvo de uma política é sua população objetivo, isto é, aquela que se enquadra nos critérios de focalização. Essas definições são discutidas no Capítulo 3 do volume "A política é nova? Avaliação ***ex ante***!" deste Guia.

compreender o contexto externo à política que pode influenciar sua implementação. Por exemplo, problemas de implementação podem surgir quando, devido a dificuldades fiscais enfrentadas por algum município, ocorrem atrasos nos pagamentos dos fornecedores de um programa, os quais, por sua vez, podem deixar de prestar os serviços contratados. Assim, de forma a construir essa visão ampla acerca do objeto da avaliação de processos, é importante a participação dos formuladores e gestores da política, dos profissionais que a operam e mesmo dos seus beneficiários.

As seções a seguir abordam as etapas necessárias para a elaboração de uma avaliação de processos, seguindo especialmente as recomendações de Rossi et al. (2003) e Saunders et al. (2005). Ao final do capítulo, são apresentados dois exemplos resumidos de avaliações reais, destacando o tipo de perguntas que podem ser respondidas por avaliações de processos e as maneiras como isso pode ser feito.

3.1. Definição do período e das perguntas de avaliação

Para cumprir satisfatoriamente os objetivos de uma avaliação de processos, deve-se ter clareza sobre a etapa da política em que a avaliação será realizada. Além disso, de forma relacionada, devem ser definidas as questões a serem respondidas, tendo em vista o uso que se deseja fazer das conclusões obtidas a partir da avaliação.

Em relação à etapa da política, a avaliação pode ser realizada tanto nas fases iniciais de implementação quanto em estágios mais avançados do seu desenvolvimento. A escolha da etapa em que a avaliação será realizada deve estar de acordo com o uso pretendido dos seus resultados. Avaliações de processo realizadas em fases iniciais da política permitem verificar se a implementação foi capaz de traduzir em prática a teoria que a fundamenta. Isso permite ao gestor, por exemplo, orientar eventuais ajustes na operação a fim de se alcançar os objetivos esperados. Já avaliações de

processo realizadas em fases mais avançadas da política têm a utilidade de esclarecer os mecanismos que levam aos resultados obtidos, ou podem ainda servir na identificação de melhores práticas, seja através da comparação de diferentes experiências ou a partir da análise de diferentes processos de uma mesma política.

Além disso, devem ser definidas as questões relevantes para a avaliação. Como mencionado, as perguntas que podem ser respondidas a partir de uma avaliação de processos são aquelas relacionadas à operação da política, usualmente acerca da utilização dos bens e serviços oferecidos por ela ou de sua organização, com comparações do que foi planejado em relação ao que foi implementado de fato. Tomando como base o Modelo Lógico[7], a avaliação de processos deve se referir aos componentes listados sob as categorias de *insumos, atividades* e *produtos* da política. Conforme apresentado por Rossi *et al*. (2003), exemplos de perguntas relacionadas a uma avaliação de processos incluem:

- Quantas pessoas estão recebendo os serviços?
- Os usuários da política são parte do público-alvo?
- A quantidade, tipo e qualidade dos serviços entregues é adequada?
- A política está deixando de atender parte de seu público-alvo?
- A população sabe da existência da política?
- A política está executando as atividades previstas?
- A equipe envolvida na implementação é adequada (em quantidade e qualificação)?
- Os recursos (físicos, humanos e financeiros) disponíveis são adequados para as necessidades da política?
- A política está entregando os produtos previstos?
- Existem diferenças substanciais no desempenho da política entre localidades distintas?

[7] O Modelo Lógico de uma política reflete as etapas de seu planejamento: insumos, atividades, produtos, resultados e impactos. Mais informações podem ser encontradas no Capítulo 3 do volume "A política é nova? Avaliação ex ante!" deste Guia.

- Os usuários da política estão satisfeitos com os serviços recebidos?

É importante também identificar quais informações serão necessárias para responder às perguntas de avaliação definidas, conforme apresentado na próxima seção. A decisão sobre quais questões serão respondidas afeta diretamente os métodos necessários para coletar essas informações. Assim, faz parte do planejamento da avaliação de processos adequar essas decisões aos recursos disponíveis (tempo, pessoal, orçamento etc.). O trabalho de definição de questões e métodos é, nesse sentido, um processo iterativo, desenvolvido até que se encontre a melhor combinação de questões e métodos adequados aos recursos disponíveis.

3.2. Coleta de dados

Uma vez definidas as perguntas de avaliação, é preciso identificar: (i) quais informações são necessárias para respondê-las; (ii) que métodos são necessários para o levantamento dessas informações.

Os dados utilizados para a realização de uma avaliação de processos podem ser quantitativos ou qualitativos (HM TREASURY, 2011). Para isso, dados provenientes do sistema de monitoramento da política são de extrema importância, por fornecerem informações históricas com periodicidade definida dos indicadores relacionados à política, permitindo analisar a evolução das dimensões de interesse para a avaliação. Além disso, um bom sistema de monitoramento tem como base o Modelo Lógico da política e, portanto, já deve abordar muitos dos tópicos necessários para que as perguntas de avaliação possam ser respondidas.[8]

Além disso, dados qualitativos podem ser aliados importantes para o entendimento e interpretação das informações adquiridas através do sistema de monitoramento. Revisar documentos relacionados à contratação dos insumos necessários para implementação das atividades da política pode, por exemplo, auxiliar na verificação de problemas com fornecedores. Entrevistas

[8] O volume 2, "Como monitorar uma política pública?", deste Guia detalha as etapas envolvidas na elaboração de um sistema de monitoramento adequado.

com a equipe implementadora ou com os beneficiários da política podem elucidar questões sobre as barreiras à implementação adequada dos bens e serviços previstos, a quantidade e qualidade desses componentes, o funcionamento das parcerias com instituições locais, entre outras.

A coleta de informações também é essencial quando os dados de monitoramento não são suficientes para observar as dimensões requeridas pela avaliação de processos, devido, por exemplo, à falta de planejamento para a criação do sistema de monitoramento ou à natureza atípica de uma determinada pergunta de interesse para a avaliação. Em todos os casos, é papel da avaliação de processos coletar os dados necessários para que o resultado da avaliação seja o mais informativo possível. Os Boxes A, B, C (disponíveis no capítulo 2 deste volume) e D (a seguir) apresentam recomendações gerais sobre métodos de coleta de dados a partir de revisão de documentos, condução de entrevistas, grupos focais e aplicação de questionários, respectivamente.

Boxe D

Coleta de dados: Questionários

Um questionário é um conjunto de perguntas elaboradas para coletar informações de indivíduos, empresas etc. A aplicação pode ser feita, por exemplo, por telefone, entrevistas presenciais ou de forma eletrônica (questionários *online*).

Quando utilizar questionários para avaliação?

Utilizam-se questionários para avaliação quando os recursos são limitados e precisa-se de muitas observações, uma vez que se trata de uma alternativa relativamente barata. Por exemplo, para coletar dados sobre conhecimentos, crenças, atitudes e comportamentos. Além disso, questionários podem ser úteis quando é importante proteger a privacidade dos participantes, pois as respostas podem ser anônimas ou confidenciais.

Como planejar e desenvolver um questionário?

Definir os objetivos é a parte mais importante do questionário, para que sejam coletadas apenas informações realmente úteis à análise pretendida. Da mesma forma, deve-se selecionar a quantidade e o tipo de participantes para o questionário, tendo em vista a representatividade da amostra escolhida.

Quanto às questões incluídas, é essencial que comuniquem claramente o que se pretende saber, com palavras simples e escrita clara, evitando abreviações, jargões ou frases coloquiais. Além disso, deve-se decidir quando usar perguntas fechadas e abertas. Perguntas fechadas incluem uma lista predeterminada de respostas a partir das quais os participantes podem escolher. Perguntas abertas podem ser úteis para quando não se sabe as possíveis respostas às perguntas, para coletar *insights* ou para quando se deseja obter informações mais complexas.

É tipicamente importante incluir perguntas sobre características demográficas como sexo, raça, idade, educação, onde o participante trabalha ou reside, com o objetivo de descrever os subgrupos de entrevistados. A ordem das questões no questionário também deve ser bem planejada. Por exemplo, questões sobre as características demográficas dos indivíduos podem ser agrupadas em um mesmo "bloco" de questões. O mesmo vale para outros tipos de perguntas, como aquelas relacionadas à educação ou condições de trabalho das pessoas. Além disso, quando se deseja coletar informações mais sensíveis ou confidenciais, pode ser interessante começar com perguntas mais simples e concluir com perguntas possivelmente sensíveis. Por fim, é essencial testar o questionário através de um piloto para identificar possíveis falhas em seu fluxo de questões ou mesmo na formulação das perguntas.

Como obter uma taxa de resposta adequada?

A taxa de resposta é definida pelo número de participantes que respondem ao questionário como proporção do número total de participantes incluídos na pesquisa. Para obter uma taxa de resposta alta, é importante comunicar a importância e a finalidade do questionário, pois assim os participantes estarão possivelmente mais propensos a responder.

Fonte: adaptado de US Centers for Disease Control and Prevention, "Data Collection Methods for Evaluation: Questionnaires" Evaluation ETA Evaluation Briefs, No. 14, Novembro de 2008. Disponível em: <www.cdc.gov/healthyyouth/evaluation/pdf/brief14.pdf>.

3.3. Análise das informações coletadas

Finalmente, uma vez definida a fase da política em que a avaliação de processos deve ser conduzida, estabelecidas as perguntas de interesse a serem respondidas e coletados os dados necessários para obter essas respostas, deve ser realizada a análise dos processos operacionais da política, tomando como base principal o Modelo Lógico da mesma.

O Modelo Lógico servirá como um guia para orientar a comparação entre a implementação observada - a partir dos dados coletados para a avaliação - e o desenho inicial da política, com o propósito de entender se os processos são geridos e coordenados de modo a contribuir para a efetividade da política em alcançar os objetivos estabelecidos. Para que a lógica do programa se cumpra, é necessário que os insumos tenham sido aplicados adequadamente para a realização das atividades listadas e entrega dos produtos previstos.

Por refletir a chamada Teoria do Programa, o Modelo Lógico serve como base para identificar quais são os aspectos mais importantes relacionados ao desempenho operacional e, portanto, permite que se chegue a conclusões sobre o que pode ser considerada uma performance razoável para cada dimensão analisada (ROSSI et al., 2003). Assim, é importante que a avaliação de processos compreenda se todos os produtos foram gerados a partir das atividades propostas e se atingiram os objetivos previstos.

Em relação aos bens e serviços oferecidos, a avaliação deve prover informações acerca da capacidade operacional da política em levar esses benefícios ao público-alvo. Em última análise, eles são o elo fundamental entre o Modelo Lógico desenvolvido pelos formuladores da política e os resultados que se pretende obter. Além disso, é necessário discutir em que medida a entrega desses bens ou serviços é traduzida em uso ou exposição efetiva dos beneficiários. Esse é um ponto crítico para a verificação do desempenho, uma vez que um serviço à disposição não utilizado pelo público-alvo representa desperdício de recursos.

Em suma, uma avaliação de processos discute ao menos quatro componentes (SAUNDERS et al., 2005):

- **Aderência (*fidelity*):** comparação entre a implementação realizada e o que foi planejado inicialmente, incluindo considerações sobre a qualidade dessa entrega. Por exemplo, no caso de uma política de treinamentos, possíveis perguntas incluem se os professores cobriram todos os conteúdos de aula previstos e se os métodos de ensino foram adequados;

- **Entrega (*dose delivered*):** nível de entrega dos bens e serviços pertinentes à intervenção. Por exemplo, ainda no contexto de treinamentos, pode-se perguntar se todas as aulas de fato aconteceram (se os professores não faltaram) e se os materiais de aula foram preparados em tempo (por exemplo, elaboração de manuais para os professores ou instalação de *softwares* específicos nos computadores);

- **Exposição (*dose received*):** grau de exposição dos participantes em relação aos benefícios entregues pela política, incluindo questões como a utilização dos serviços, receptividade, nível de interação com a equipe implementadora e a satisfação com a política. Exemplos de perguntas relacionadas à exposição incluem o nível de participação dos alunos em aula (por exemplo, interações dos alunos com os professores) e nível de utilização dos materiais de aula (por exemplo, se os professores baseavam suas aulas em manuais previamente elaborados para isso);

- **Alcance (*reach*):** nível de atendimento do público-alvo pretendido, incluindo considerações sobre a proporção do público-alvo que de fato é beneficiada pela política (grau de *cobertura*[9]). Por exemplo, se um programa consiste no ensino extracurricular de língua inglesa aos alunos da oitava série de escolas estaduais, pode-se perguntar qual a porcentagem desse público-alvo que de fato participou dos cursos oferecidos.

[9] O grau de cobertura é o tamanho da população que está de fato sendo atendida (população beneficiária) em relação ao tamanho da população que se pretende atender (população objetivo), expresso em termos percentuais. Esses conceitos são definidos no Capítulo 3 do volume "A política é nova? Avaliação *ex ante*!" deste Guia.

Em todos esses casos, é importante contrastar os valores encontrados com as metas estabelecidas inicialmente. Só será possível julgar a adequação da implementação realizada quando ela for comparada a algum valor de referência (as próprias metas, por exemplo). Além disso, deve-se discutir em quais situações e em que medida as diretrizes da política (por exemplo, procedimentos burocráticos para contratação de fornecedores) entravam o desenvolvimento dos processos, tornando ineficazes os componentes operacionais. Essa é uma oportunidade de se apontar sugestões de melhoria ou adequação dessas diretrizes.

Seguindo o modelo de México (2017), propostas de alterações na implementação da política devem considerar: (i) a viabilidade de implementação da proposta apresentada; (ii) os agentes que seriam responsáveis por essa implementação; (iii) os potenciais efeitos que a mudança proposta poderia ter, tanto sobre a operação da política quanto sobre o potencial de alcance de seus objetivos, levando em consideração as possíveis restrições de ordem prática; e (iv) uma comparação entre a situação que se observou e o que se esperaria obter uma vez implementada a recomendação.

Exemplo 1

Dois programas educacionais

No relatório "*A wide angle view of learning Evaluation of the CCE and LEP programmes in Haryana, India*", Duflo et al. (2015) avaliam os impactos dos programas indianos *Continuous and Comprehensive Evaluation* (CCE) e *Learning Enhancement Program* (LEP) sobre o indicadores de desempenho dos alunos de escolas públicas do estado de Haryana, na Índia, utilizando o método experimental[10]. Ao longo do relatório, também são apresentados os resultados de uma avaliação dos processos de ambos os programas, realizada após preocupações terem sido levantadas quanto à correta implementação dos mesmos.

[10] Mais informações sobre a utilização do método experimental para realizar avaliação de impacto podem ser encontradas no Capítulo 4 deste volume do Guia.

Intervenção avaliada

Conforme apresentado por Duflo et al. (2015), o sistema CCE consiste em avaliar regularmente os alunos a partir de notas atribuídas não apenas com base no desempenho escolar, mas também em atividades extracurriculares (como artes, música ou atletismo) e no desenvolvimento da personalidade (conforme refletido nas habilidades, atitudes e valores), permitindo que os professores personalizem seu método de ensino conforme os níveis de aprendizagem de cada aluno. Já o LEP baseia-se na ideia de "ensinar no nível certo" (TaRL, "*teaching at the right level*"), fornecendo ferramentas e alocando tempo no horário escolar para que os professores consigam adequar o nível de ensino às habilidades de cada criança. Ambos os programas pretendem abordar o baixo desempenho dos alunos refletido, principalmente, pela dificuldade em alcançar o desempenho necessário em cada série.

Foram selecionadas aleatoriamente 500 escolas nos distritos de Mahendragarh e Kurukshetra para receber o tratamento durante o ano letivo de 2012-2013. Além de participar de treinamentos – incluindo recomendações sobre como conduzir avaliações regulares e manter registros do progresso de cada aluno –, os professores também receberam materiais (como manuais, fichas de avaliação, boletins individuais, entre outros) para implementar os programas. O desempenho dos alunos foi medido por avaliações orais e escritas (habilidades básicas de Hindi e Matemática).

Metodologia da avaliação de processos

Para avaliar os processos do CCE e do LEP, foi estabelecido um programa de monitoramento de processos constituído por duas visitas surpresas, entre agosto de 2012 e março 2013, a cada uma das 500 escolas incluídas na avaliação. Durante as visitas, os entrevistadores eram responsáveis por aplicar um amplo questionário aos professores, abordando questões sobre a implementação dos programas, incluindo perguntas relacionada à utilização dos insumos de aprendizagem (livros didáticos, uniformes, manuais, entre outros). Os entrevistadores também selecionaram aleatoriamente um professor em cada escola para coletar informações sobre as práticas de ensino e avaliação em sala de aula.

Principais resultados da avaliação de processos

Duflo et al. (2015) reportam os seguintes resultados em relação aos processos do CCE e do LEP:

- *Escolas selecionadas para receber o CCE*: 75% dos professores participaram do treinamento. Na visita final de monitoramento, 88,7% dos professores que receberam o treinamento relataram ter seus manuais, e apenas 42,2% conseguiram mostrá-lo ao entrevistador. Sobre o uso dos materiais distribuídos pelos programas, 81,7% e 64,7% dos professores relataram utilizar as fichas de avaliação e os boletins individuais, respectivamente. Entretanto, apenas 45,2% conseguiram mostrar as fichas de avaliação ao entrevistador (38,6% para os boletins individuais).
- *Escolas selecionadas para receber o LEP*: 93,9% dos professores participaram do treinamento. Na visita final de monitoramento, 98,1% dos professores que receberam o treinamento relataram ter seus manuais, mas somente 64,6% conseguiram mostrá-lo ao entrevistador. Sobre o uso de fichas de avaliação, 85% das escolas haviam preenchido, e 99,4% dos diretores informaram que conduziram aulas de reforço todos os dias.

Exemplo 2

Programa de proteção social

No relatório "*Can e-governance reduce capture of public programmes? Experimental evidence from India's employment guarantee scheme in Bihar*", Banerjee et al. (2015) avaliam os impactos de uma reforma no sistema de pagamentos do programa de proteção social *Mahatma Gandhi National Rural Employment Guarantee Scheme* (MGNREGS) sobre indicadores de gastos do programa, nível de emprego e de salários. No relatório, também são apresentados os resultados de uma avaliação de processos, realizada para clarificar os motivos das baixas taxas de adesão à reforma proposta.

Intervenção avaliada

Conforme apresentado em Banerjee et al. (2015), o MGNREGS é um dos maiores programas de proteção social do mundo e contava, em 2013, com cerca de 50 milhões de famílias beneficiárias. O programa consiste em ofertar 100 dias de trabalho por ano para adultos em zonas rurais da

Índia que aceitem as condições de salários e as ocupações disponíveis. Bihar, possivelmente o mais pobre entre os grandes estados da Índia, tem a menor taxa de participação no MGNREGS entre todos os estados. Segundo os autores, essa baixa participação ocorre por duas razões principais: a falta de capacidade administrativa e a corrupção, que limitam o desempenho do programa.

A estrutura burocrática do MGNREGS tornava necessária a intermediação dos recursos financeiros do programa (utilizados para pagar os salários dos beneficiários) por diversos agentes, organizados em subdivisões regionais. Buscando diminuir os problemas do programa em termos de capacidade administrativa e corrupção (possivelmente originados de sua estrutura altamente burocrática), uma reforma foi proposta. Essa reforma consistia em possibilitar que os agentes locais acessassem diretamente o *software* financeiro por onde as requisições de pagamentos deveriam ocorrer, reduzindo, portanto, o número de agentes envolvidos nesse processo. A reforma foi implementada em 69 blocos (subdivisões regionais) escolhidos aleatoriamente em 12 distritos de Bihar, nos anos de 2012 e 2013. Em cada bloco foi realizado um treinamento sobre a nova estrutura do sistema de pagamentos.

Metodologia da avaliação de processos

Para avaliar os processos relacionados à reforma do sistema de pagamentos do MGNREGS, foram realizadas entrevistas com os agentes locais do programa. Além disso, foram consultados dados de monitoramento provenientes de planilhas de acompanhamento da disponibilidade de recursos dos blocos, assim como informações de um portal de reclamações dos agentes sobre o *software* financeiro, desenvolvido para agilizar o processo de resolução de problemas de infraestrutura.

Principais resultados da avaliação de processos

Banerjee et al. (2015) reportam os seguintes resultados em relação aos processos da intervenção avaliada:

- *Primeira fase de implementação (2012)*: menos de 20% dos agentes locais utilizaram o novo sistema. Verificou-se que essa baixa taxa de utilização foi consequência da falta de recursos estaduais para o funcionamento do MGNREGS, que acarretou no atraso dos salários dos agentes locais e, em última instância, em uma greve geral desses agentes. Apesar de a greve não ter sido causada diretamente pela reforma do sistema de pagamentos, os funcionários que tiveram salários atrasados associaram os

atrasos à reforma, o que colaborou para a baixa taxa de utilização do novo sistema e, portanto, atrasou a implementação.

- *Segunda fase de implementação (2013)*: em média, 60% dos agentes locais adotaram o novo sistema de pagamentos durante a segunda fase. Além disso, a qualidade da implementação variou nos distritos. Em Begusarai, por exemplo, o acesso ao sistema passou de 20% para 90% entre as duas fases da implementação, enquanto em Madhubani, não ultrapassou os 40%.

Dois problemas principais foram identificados como causas dessa falha de implementação. Primeiro, a reforma causou um aumento do número de pagamentos a serem processados pelo Banco Central da Índia. O Banco, por sua vez, não possuía capacidades administrativas para garantir o processamento de todos os pagamentos, o que culminou na ocorrência de novos atrasos nos salários dos agentes e diminuiu a adesão à reforma. Em segundo lugar, a reforma tornava os agentes locais responsáveis por registrarem, junto aos sistemas de informação do Ministério do Desenvolvimento Rural da Índia, cada pagamento de salários aos beneficiários do MGNREGS. Esses registros deveriam ocorrer automaticamente após os agentes preencherem as informações necessárias no *software* financeiro do próprio programa. Entretanto, por falta de coordenação entre o Ministério das Finanças e o Ministério do Desenvolvimento Rural, a integração entre o *software* financeiro e os demais sistemas de informação não foi desenvolvida. Essa falta de integração implicava na necessidade de documentar cada pagamento duas vezes, o que sobrecarregaria os agentes locais e, portanto, também afetou negativamente a adesão à reforma do MGNREGS.

4 Avaliação de impacto

O termo *avaliação de impacto* diz respeito a uma *abordagem quantitativa de avaliação*, que tem como objetivo aferir os impactos gerados a partir de uma política pública ou intervenção de interesse. Essas avaliações procuram identificar *relações causais*, isto é, de causa e efeito, entre a política pública avaliada e as mudanças observadas nas dimensões de interesse. Especificamente, esse tipo de avaliação pretende verificar se a intervenção avaliada teve ou não impactos sobre determinadas dimensões de interesse e, caso tenha tido, quantificar os impactos médios observados, provendo resultados como: aumento de 2 pontos nas notas de Língua Portuguesa dos alunos do 5º ano do Ensino Fundamental, incremento de 245 reais mensais na renda familiar, redução em 12% dos casos de incidência de diabetes.

Nesse sentindo, a avaliação de impacto é fundamental para avaliar se a política pública analisada é capaz de promover as transformações pretendidas originalmente e desempenha um papel-chave no contexto de políticas baseadas em evidências.

4.1. Protocolo de avaliação de impacto

Para identificar *relações causais* entre a política e as mudanças observadas, as avaliações de impacto são sempre baseadas em *comparações* entre: (i) o que se observa na realidade, uma vez ocorrida a política pública; e (ii) o que se espera que se observaria caso a política em questão não tivesse sido adotada – a situação *contrafactual*. A intuição desse método é que, ao comparar as situações (i) e (ii) em um momento posterior à implementação da política avaliada, quaisquer diferenças observadas entre elas no que diz respeito às dimensões de impacto de interesse (ex: educação, saúde) poderá ser considerada fruto do acesso à intervenção, uma vez que essa é a única diferença prévia entre as duas situações. Sabe-se, no entanto, que não é possível observar o contrafactual na prática, já que não se pode observar o mesmo grupo (beneficiários da política pública avaliada) em duas situações mutua-

mente excludentes (tendo participado vs. não tendo participado) em um mesmo momento de tempo (pós-intervenção)[11].

Para que se possa conduzir uma avaliação de impacto, portanto, será preciso estimar o contrafactual. Isso é feito a partir da seleção de um *grupo de comparação* e da observação dos resultados desse grupo nas dimensões de interesse.

Boxe E

Resultados potenciais

Formalmente, seja T_i uma variável binária indicativa de tratamento do indivíduo *i* pela política em questão, tal que $T_i = 1$ se o indivíduo foi tratado ($T_i = 0$) se o indivíduo *i* não foi tratado. Sejam ainda Y_i^1 e Y_i^0 os resultados potenciais para o indivíduo *i* quando o mesmo é tratado ($T_i = 1$) ou não tratado ($T_i = 0$), respectivamente. O impacto do tratamento (ser beneficiado pela política em questão) para o indivíduo i é dado por $\beta_i = Y_i^1 - Y_i^0$. No entanto, se o indivíduo foi tratado ($T_i = 1$), observa-se apenas Y_i^1 ao passo que se o mesmo não foi tratado ($T_i = 0$) observa-se apenas Y_i^0. Dessa forma, tem-se que o resultado de fato observado na prática, Y_i, é dado por: $Y_i = Y_i^1 T_i + Y_i^0 (1-T_i)$.

4.1.1. Grupos de tratamento e de comparação

Em uma avaliação de impacto, comparam-se os grupos:

- **Grupo de tratamento:** composto por indivíduos ou unidades (por exemplo, escolas, bairros, postos de saúde etc.) que foram beneficiados pela política pública em análise.
- **Grupo de comparação:** composto por indivíduos ou unidades que não foram beneficiados pela política pública em análise, mas que apresentam características similares àquelas dos beneficiados pela mesma. Esse grupo é comumente denomi-

[11] Essa questão é apresentada em detalhes e discutida em Rubin (1974).

nado *grupo de controle*[12] e deve representar o que teria acontecido com o grupo de tratamento se ele não tivesse sido beneficiado pela política (isto é, a situação contrafactual).

Quanto mais parecidos forem os grupos de tratamento e de comparação, maior será a garantia de que, ao compará-los, será possível identificar o efeito causal de ter sido beneficiado pela política pública em análise. Conforme discutido em Gertler et al. (2018), é importante que o grupo de tratamento e o grupo de comparação sejam semelhantes no que diz respeito às suas características médias na ausência da intervenção, ao contexto no qual estão inseridos e à forma como reagem ao tratamento.

Caso o grupo de tratamento e o grupo de comparação apresentem *outras diferenças* para além do fato de o primeiro ter sido beneficiado pela política em análise enquanto o segundo não, a comparação dos resultados dos mesmos nos dará uma estimativa *viesada* do impacto da intervenção. Isso porque quaisquer diferenças observadas entre os resultados futuros dos mesmos serão compostas pelo impacto acrescido dos efeitos provenientes dessas outras diferenças.

Suponha, por exemplo, uma política pública de assistência social para famílias de baixa renda cuja participação dependa da inscrição voluntária por parte das famílias. Para avaliar o impacto dessa política sobre o desempenho em um indicador de pobreza, considere que foi proposto comparar esse resultado entre as famílias beneficiadas (grupo de tratamento) e famílias não beneficiadas (grupo de comparação). Nesse caso, o chamado *viés de seleção* pode refletir o fato de que as famílias beneficiárias são diferentes daquelas famílias do grupo de comparação, por exemplo, em termos de motivação ou habilidades não observadas, refletidos justamente na decisão dessas famílias de realizarem a inscrição no programa.

[12] O termo "grupo de controle" é utilizado quando a metodologia de avaliação de impacto é experimental.

Boxe F

Viés de seleção

A diferença observada no indicador de impacto Y_i entre os grupos de tratamento ($T_i=1$) e comparação ($T_i=0$) pode ser representada da seguinte forma:

$$Diferença = E[Y_i|T_i=1] - E[Y_i|T_i=0] = E[Y_i^1|T_i=1] - E[Y_i^0|T_i=0]$$

em que $E[Y_i|T_i=1]$ corresponde ao valor esperado do indicador para os indivíduos ou unidades do grupo de tratamento (por exemplo, beneficiários de uma política) e $E[Y_i|T_i=0]$ ao mesmo para o grupo de comparação proposto. Ao somar e subtrair o termo $E[Y_i^0|T_i=1]$ da expressão anterior, tem-se que:

$$Diferença = E[Y_i^1|T_i=1] - E[Y_i^0|T_i=0] + E[Y_i^0|T_i=1] - E[Y_i^0|T_i=1]$$

$$= \underbrace{(E[Y_i^1|T_i=1] - E[Y_i^0|T_i=1])}_{\text{Impacto}} + \underbrace{(E[Y_i^0|T_i=1] - E[Y_i^0|T_i=0])}_{\text{Viés de seleção}}$$

Dessa forma, a seleção de um grupo de comparação adequado é fundamental para que a avaliação de impacto seja capaz de mensurar corretamente os impactos gerados pela política em análise. Um grupo de comparação adequado será aquele que permitirá que se elimine eventuais vieses de seleção, tal que a diferença entre os grupos corresponda apenas ao impacto. No exemplo anterior da política pública de assistência social, a *autosseleção* via inscrição voluntária por parte das famílias é um exemplo de possível fonte de viés.

Para selecionar um grupo de comparação adequado, é preciso compreender os critérios de elegibilidade e de priorização da política em análise, uma vez que eles definirão as características que devem ser equalizadas entre os grupos, possíveis candidatos a grupos de comparação, além de terem papel determinante na escolha da metodologia da avaliação, conforme discutido a seguir.

4.2. Metodologias de avaliação de impacto

As metodologias de avaliação de impacto podem ser divididas entre os chamados métodos experimentais e métodos não experimentais. O método experimental pode ser utilizado no caso específico em que a seleção de beneficiários da política pública em análise se dá a partir de um processo de aleatorização (sorteio). Já os métodos não experimentais são utilizados quando os critérios para a seleção de beneficiários tomam outras formas, como seleção com base em características observáveis (ex: renda, condição de emprego, escolaridade) ou desempenho em um indicador específico para o qual exista um ponto de corte bem definido (ex: desempenho em exame de admissão).

Conforme será discutido a seguir, esses métodos diferem entre si no que diz respeito às hipóteses nas quais se baseiam, ao grupo de comparação a ser utilizado e à forma como de fato são comparados os grupos de tratamento e comparação.

4.2.1. Método experimental

O chamado método experimental de avaliação de impacto é utilizado nos casos em que o status de tratamento, isto é, ser beneficiado pela política pública em análise, é atribuído de maneira aleatória aos indivíduos ou unidades elegíveis. Algumas aplicações desse método são os *randomized controlled trials* (experimentos aleatórios), que são utilizados, por exemplo, na área de saúde para testar novos medicamentos.

Do ponto de vista de implementação, a seleção de beneficiários por meio de um processo de aleatorização envolve primeiramente a listagem de todos os indivíduos ou unidades elegíveis à política. Nesse sentido, é fundamental que os critérios de elegibilidade sejam bem definidos e objetivos. Alguns exemplos de possíveis critérios de elegibilidade seriam: escolaridade mínima de ensino médio completo; renda familiar máxima igual a 1 salário mínimo; residência nos municí-

pios A, B ou C; inscrição para participação até uma data limite. Uma vez definida a lista de elegíveis, o passo seguinte é atribuir de forma aleatória a cada unidade elegível seu status de tratamento pela política: selecionado ou não selecionado. Esse processo pode ser feito de diferentes formas, sendo uma delas atribuir um número aleatório para cada unidade elegível, ordená-las a partir desse número e definir que as unidades nas primeiras posições serão selecionadas para serem beneficiadas pela política pública em questão, até que sejam preenchidas todas as vagas ou extintos os recursos disponíveis. É importante ressaltar que todo esse processo deverá ser planejado durante a etapa de desenho da política, já que fará parte do início da implementação.

Considerando a credibilidade e robustez da avaliação de impacto, esse método é o preferível, uma vez que, ao atribuir de forma aleatória o status de tratamento entre as unidades elegíveis, é possível garantir que, em média, os grupos formados, de selecionados (tratamento) e não selecionados (comparação), apresentarão características pré-intervenção similares, tanto no que se refere à aspectos observáveis (ex: idade, renda) quanto não observáveis (ex: interesse, engajamento).

Essa homogeneidade entre os grupos, no entanto, terá alta probabilidade de ocorrer apenas se o número de unidades elegíveis e o tamanho dos grupos forem suficientemente grandes, conforme será discutido posteriormente. Para características observáveis, caso haja dados de períodos pré-intervenção disponíveis para ambos os grupos, é importante conduzir testes de balanceamento entre os grupos, isto é, testar se estes de fato são similares em termos das médias observadas para essas características. Caso haja indícios de que há desbalanceamento, será necessário repensar o uso do método experimental, utilizando-se métodos não experimentais ou pelo menos controlando pelas características para as quais foram encontradas diferenças prévias entre os grupos.

Intuitivamente, ao garantir a similaridade pré-intervenção entre os grupos de tratamento e de comparação, elimina-se o viés de seleção[13], de forma que quaisquer diferenças posteriores observadas entre eles poderão ser consideradas como sendo justamente o impacto de ter sido beneficiado pela política em questão. mesma. Esse grupo é comumente denomi

Boxe G

Impacto

Formalmente, a aleatorização do status de tratamento implica que a designação do tratamento será independente dos resultados potenciais:

$$(Y_i^1, Y_i^0) \perp T_i \rightarrow E[Y_i^0 | T_i=1] = E[Y_i^0 | T_i=0] = E[Y_i^0]$$

Dessa forma, o viés de seleção é igual a zero e a diferença observada é igual ao impacto:

$$Diferença = (E[Y_i^1 | T_i=1] - E[Y_i^0 | T_i=1]) = Impacto$$

Boxe H

Estimação do impacto utilizando o método experimental

Nesse método, a estimação do impacto da política pode ser feita a partir de uma regressão linear simples:

$$Y_i = \alpha + \beta T_i + e_i$$

em que Y_i corresponde ao indicador de impacto de interesse para o indivíduo i, T_i é uma variável binária indicativa de tratamento (assume o valor *1* para indivíduos do grupo de tratamento e o valor *0* para indivíduos do grupo de comparação) e e_i é o termo de erro. O impacto da política será dado por β, lembrando que além de estimá-lo, deve-se conduzir um teste de hipóteses para verificar sua significância estatística (isto é, verificar se o impacto estimado é estatisticamente diferente de *0*).

[13] Esse aspecto é discutido em maiores detalhes em Angrist e Pischke (2008, Capítulo 2) e também em Menezes Filho e Pinto (2017, Capítulo 3).

Além da vantagem do ponto de vista da robustez da avaliação de impacto, a seleção de beneficiários por meio de um processo de aleatorização pode também ser vantajosa ao se considerar a facilidade de implementação, transparência e igualdade de acesso associadas a ele, sobretudo quando há excesso de demanda (isto é, o número de unidades elegíveis à política pública excede o número total de unidades que poderão ser atendidas tendo em vista o número de vagas ou de recursos disponíveis)[14]. No entanto, pode haver casos em que esse tipo de seleção não é desejável ou factível por razões éticas ou legais, por exemplo. Dessa forma, é preciso analisar de forma criteriosa a aplicação desse tipo de seleção e de avaliação quando se consideram novos projetos de políticas públicas.

Exemplo 1

PROGRESA - México

No artigo "*Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment*", Gertler (2004) utiliza o método experimental para avaliar os impactos do programa mexicano PROGRESA sobre indicadores de saúde infantil.

Intervenção avaliada

Conforme apresentado em Gertler (2004), o PROGRESA é um programa mexicano de transferência condicional de renda que se iniciou em 1997 e tem como objetivo combater a pobreza, considerando inclusive aspectos relacionados à transmissão intergeracional da mesma. A seleção para participação no programa se dá a partir da escolha de comunidades consideradas em situação de vulnerabilidade[15] e posterior seleção das famílias de baixa renda dentro dessas comunidades. O programa apresenta diversas condições para o recebimento da transferência de renda, destacando-se: (i) no caso de crianças de até cinco anos, exigência de imunização e de visitas regulares a clínicas de monitoramento da nutrição infantil, onde são distribuídos suplementos nutricionais e são realizadas atividades educativas com os pais sobre nutrição, saúde e higiene; (ii) no caso de gestantes e lactantes, exigência de visitas a

[14] Gertler ***et al.*** (2018) discutem esses aspectos em maiores detalhes.

[15] A vulnerabilidade era medida a partir de um índice específico que contemplava aspectos relacionados à escolaridade e tipo de trabalho dos adultos e características dos domicílios (ex: acesso à água e eletricidade, número médio de moradores por cômodo), conforme apresentado em Skoufias et al (1999).

clínicas de saúde para consultas de pré-natal ou pós-parto, recebimento de suplementos alimentares e participação em atividades educativas sobre saúde; (iii) no caso de outros adultos, exigência de visitas a clínicas de saúde para consultas preventivas (*check-ups*) e para participar de atividades educativas sobre saúde.

Metodologia de avaliação de impacto

Para avaliar os impactos do PROGRESA sobre os indicadores de saúde infantil de altura, incidência de anemia e de morbidade[16], Gertler (2004) faz uso do desenho experimental (*randomized controlled trial - RCT*) utilizado pelo governo mexicano para a seleção para participação no ano de 1998. Devido a restrições orçamentárias, não era possível atender a todas as vilas elegíveis naquele momento, de forma que decidiu-se sortear as vilas que receberiam o programa primeiro. No total, foram sorteadas 320 vilas para receber o programa naquele ano (grupo de tratamento) e 185 para receber o programa posteriormente (grupo de controle). Foi realizada uma pesquisa de campo para a coleta de informações e a amostra selecionada para as análises foi restrita às famílias consideradas elegíveis para receber o programa, segundo o critério de renda, em ambos os grupos (tratamento e controle). Os impactos foram estimados a partir da comparação das médias dos indicadores de interesse entre os grupos[17], controlando por características socioeconômicas consideradas relevantes para aumentar a precisão das estimativas (ex: sexo e idade da criança, idade e escolaridade dos pais, renda familiar em um momento pré-intervenção).

Principais resultados da avaliação de impacto

Gertler (2004) reporta os seguintes impactos do PROGRESA:

- *Altura*: em média, crianças do grupo de tratamento cresceram 1 centímetro a mais que as crianças do grupo de controle durante o primeiro ano do programa.
- *Anemia*: em média, crianças do grupo de tratamento tinham probabilidade menor de estarem anêmicas durante o primeiro ano do programa (25,3%).
- *Morbidade*: crianças de famílias do grupo de tratamento tiveram, em média, incidência de doenças menor que crianças de famílias do grupo de controle (25,3% para crianças de até seis meses de idade nascidas durante a intervenção e 22,3% para crianças de até três anos).

[16] O indicador para morbidade utilizado corresponde a uma variável binária que indica se a mãe reportou que a criança tinha ficado doente nas quatro semanas anteriores à pesquisa.

[17] Para realizar a comparação de médias, foi utilizada regressão linear no caso de altura e regressões logísticas para os demais indicadores. Regressões logísticas são utilizadas quando se deseja estimar um modelo em que a variável dependente é uma variável do tipo categórica (ex: variável binária indicativa de anemia). Nesse caso, considera-se que esta segue uma distribuição logística.

4.2.2. Métodos não experimentais

Os métodos não experimentais de avaliação de impacto são utilizados quando a seleção para participação em uma intervenção baseia-se em características observáveis ou não observáveis. A seguir, serão discutidos os métodos de *pareamento* para o caso de seleção em observáveis e de *diferença em diferenças, variável instrumental* e *regressão descontínua* para casos de seleção em não observáveis.

Pareamento

O método de pareamento (*matching*) é utilizado quando a seleção para participação na intervenção em análise baseia-se em características observáveis, exclusivamente. A intuição do método é que, condicional a um conjunto de características observáveis, ser tratado ou não pela política em análise pode ser considerado como sendo aleatório.

Nesse método, selecionam-se dentre as unidades não beneficiadas pela política aquelas que apresentam características observáveis pré-intervenção mais semelhantes às das unidades beneficiadas para compor o grupo de comparação. Essas unidades não beneficiadas selecionadas são utilizadas para estimar o contrafactual, ou seja, aquilo que teria acontecido com as unidades beneficiadas na ausência da política. Na prática, para cada unidade do grupo de tratamento (composto pelos beneficiários da política pública em questão), selecionam-se para serem seus pares uma ou mais unidades não beneficiadas que sejam as mais parecidas possíveis, utilizando-se técnicas econométricas específicas e levando em consideração as características observáveis que são relevantes para determinar a participação na intervenção e as expectativas sobre o futuro.

Boxe I

Hipóteses do método de pareamento

O método de pareamento tem como hipóteses[18]:

1. *Independência condicional*: condicional a um conjunto X_i de variáveis observáveis, os resultados potenciais (Y_i^1, Y_i^0) são independentes do tratamento T_i, tal que $(Y_i^1, Y_i^0) \perp T_i | X_i$.

2. *Suporte comum*: sendo $p(X)$ a probabilidade de ser tratado, então $0 < p(X) < 1$. Isso significa que não há valores de X que determinem certamente o status de tratamento, o que garante a comparabilidade entre os grupos de tratamento e comparação.

Quando a quantidade de características relevantes para o processo de pareamento é elevada e/ou quando estas incluem variáveis contínuas, utiliza-se o método de pareamento com base no escore de propensão (*propensity score matching*[19]). Nesse caso, ao invés de buscar para cada unidade tratada uma unidade que apresente exatamente os mesmos valores para cada uma das características relevantes, estima-se o escore de propensão com base nessas variáveis para cada unidade tratada e não tratada. Essa medida corresponde à probabilidade estimada de ser tratado com base nas características consideradas relevantes para o pareamento, de forma que pode assumir valores entre 0 e 1. Vale ressaltar que devem ser consideradas no pareamento apenas características anteriores à intervenção, que não tenham sido afetadas pela política em análise. A seleção de quais características devem ser levadas em consideração no pareamento deve ser feita de forma criteriosa e embasada, tomando-se o cuidado de não incluir variáveis que possam ter sido afetadas, ainda que indiretamente, pela política em análise, como via expectativas de ser beneficiado, por exemplo.

[18] As hipóteses do método de Pareamento são discutidas em maiores detalhes em Menezes Filho e Pinto (2017), Capítulo 5.

[19] Rosenbaum e Rubin (1983) apresentam o método em detalhes.

A partir dos escores de propensão estimados para cada unidade (tratadas e não tratadas), no método, para cada unidade tratada serão selecionadas uma ou mais unidades não tratadas que apresentem escores de propensão os mais próximos possíveis do da primeira. É preciso definir qual critério de pareamento será utilizado, o que envolve um *trade-off* entre viés e variância nas estimações de impacto, reforçando que se deseja minimizar ambos. Por exemplo, limitar a apenas 1 par por unidade tratada reduz o viés, já que será selecionada apenas a unidade não tratada mais próxima possível. Por outro lado, utilizar vários pares para cada unidade tratada aumenta a quantidade de informação utilizada para estimar o contrafactual, de forma que se reduz a variância do estimador de impacto (ou seja, se aumenta sua precisão).

Existem diversas possibilidades de critérios de pareamento, por exemplo[20]:

- *Vizinho mais próximo*: para cada unidade tratada, seleciona-se uma ou mais unidades não tratadas que apresentem escores de propensão os mais próximos possíveis ao dela em termos da distância absoluta;
- *Raio*: define-se um valor máximo tolerável para a distância absoluta entre os escores de propensão das unidades tratada e não tratada e consideram-se como sendo pares de cada unidade tratada todas as unidades não tratadas que se encontrem dentro desse intervalo (raio);
- *Kernel*: todas as unidades não tratadas poderão ser incluídas no grupo de comparação, mas a cada uma é atribuído um peso distinto com base nas distâncias observadas entre os escores de propensão.

Para calcular o impacto de uma política pública utilizando o método de pareamento com base no escore de propensão, deve-se:

1) Estimar o escore de propensão para cada unidade (tratadas e não tratadas) a partir do uso de

[20] Caliendo e Kopeinig (2008) exemplificam esses e outros critérios de pareamento, além de apresentarem informações adicionais sobre o método.

modelos *probit* ou *logit*, em que a variável dependente corresponde à participação na intervenção e as variáveis independentes correspondem às características observáveis pré-intervenção consideradas relevantes.

2) Formar os pares de unidades tratadas e não tratadas a partir do critério de pareamento escolhido, lembrando que devem ser consideradas apenas as observações no suporte comum;

3) Estimar o impacto da política pública sobre os indicadores de impacto de interesse a partir da comparação entre: (a) o grupo de tratamento, que será composto por unidades tratadas para as quais identificou-se um ou mais pares; e (b) o grupo de comparação, que será composto apenas pelas unidades não tratadas que foram pareadas às unidades tratadas. Isso pode ser feito a partir do uso de regressões lineares incluindo-se controles e possivelmente utilizando funções dos escores de propensão como peso. Vale ressaltar que os resultados estarão restritos à região de suporte comum e correspondem a uma média das diferenças nos resultados entre os pares formados.

Embora o método de pareamento seja bastante versátil, ele só é recomendado quando a seleção para participação se baseia exclusivamente em características observáveis, vide a hipótese de independência condicional, que é bastante forte e não pode ser testada na prática. Ademais, o método requer que estejam disponíveis informações pré-tratamento e que o tamanho dos grupos seja grande para que se possam adotar os procedimentos econométricos necessários.

Exemplo 2

Programa Bolsa Família

No artigo "*Alimentação, nutrição e saúde em programas de transferência de renda: evidências para o Programa Bolsa Família* ", Camelo et al. (2009) utilizam o método de pareamento para avaliar os impactos do Programa Bolsa Família sobre indicadores de segurança alimentar das famílias e de saúde infantil.

Intervenção avaliada

Conforme apresentado em Camelo et al. (2009), o Programa Bolsa Família foi criado em 2004, unificando outros programas de transferência de renda já existentes. Trata-se de um programa de transferência condicional de renda que tem como objetivo combater a pobreza e contribuir para o rompimento do ciclo de perpetuação da mesma, tendo como foco famílias com crianças e jovens de até dezessete anos e/ou gestantes. A transferência é composta tanto por um valor fixo quanto por valores variáveis a depender da composição familiar. As condicionalidades que devem ser cumpridas incluem: (i) matrícula e frequência mínima na escola para crianças e jovens; (ii) imunização e acompanhamento médico para crianças de até seis anos; e (iii) realização de exames e acompanhamento médico para gestantes e nutrizes. O processo de seleção para o programa envolve o cadastramento realizado pelos municípios (Cadastro Único CadÚnico) e a seleção dos beneficiários feita pelo Ministério do Desenvolvimento Social, com base nas cotas de benefícios municipais disponíveis.

Metodologia de avaliação de impacto

Para avaliar os impactos do Bolsa Família sobre indicadores de segurança alimentar das famílias (Escala Brasileira de Insegurança Alimentar - EBIA) e de saúde infantil para crianças de até seis anos (altura por idade, peso por idade, peso por altura, índice de massa corporal (IMC) e mortalidade), Camelo et al. (2009) utilizam dados da Pesquisa Nacional de Demografia e Saúde (PNDS) de 2006 (Ministério da Saúde).

Os autores definem como grupo de tratamento as famílias beneficiárias do programa e selecionam as famílias não beneficiárias que seriam elegíveis ao Bolsa Família que apresentam características observáveis as mais parecidas possíveis às primeiras para compor o grupo de comparação. Para isso, utilizam o método de pareamento com base no escore de propensão (*propensity score matching*), considerando o critério de vizinho mais próximo (1 vizinho apenas). O

escore de propensão foi estimado considerando características de idade e escolaridade do chefe do domicílio, composição familiar (variável binária de família biparental, número de crianças de até seis anos e número de crianças de sete a quinze anos), variáveis de localização geográfica do domicílio (variáveis binárias para macrorregiões e para região urbana) e características de infraestrutura do domicílio (variáveis binárias de acesso à água encanada e eletricidade, número de banheiros e densidade morador-cômodo).

Principais resultados da avaliação de impacto

Camelo et al. (2009) reportam que:

- *Segurança Alimentar*: o Programa Bolsa Família aumenta, em média, em 7,4 pontos percentuais a probabilidade de o domicílio estar em situação de segurança alimentar. Ao investigar separadamente os efeitos em domicílios com diferentes graus de insegurança alimentar, somente foram encontrados efeitos do programa naqueles em situação de insegurança alimentar leve;
- *Indicadores Antropométricos*: não foram encontrados impactos estatisticamente significantes do Programa Bolsa Família sobre crianças com estado nutricional abaixo do ideal, mas foram encontrados impactos positivos na probabilidade de sair de uma situação de sobrepeso para outra de peso adequado, por idade e altura (crianças beneficiárias têm entre 5 e 7 pontos percentuais mais chances de estar com peso adequado em relação à situação de sobrepeso);
- *Mortalidade Infantil*: não foram encontrados impactos estatisticamente significantes do Programa Bolsa Família sobre esse indicador.

Diferença em Diferenças

O método de diferença em diferenças pode ser utilizado quando a seleção para participação na política pública em análise é baseada em características observáveis ou não observáveis, desde que se disponha de dados para pelo menos dois momentos no tempo: antes e depois da intervenção. O cálculo do impacto nesse caso é feito a partir da comparação das variações observadas no indicador de impacto ao longo do tempo (antes vs. depois da intervenção) para os grupos de tratamento e de comparação.

Esse método requer hipóteses mais fortes que os métodos anteriores, sendo elas[21]:

1) Na ausência da intervenção, os indicadores de impacto dos grupos de tratamento e de comparação apresentariam *tendências paralelas*. Essa hipótese implica que a variação do indicador de impacto observada ao longo do tempo para o grupo de comparação representa a variação que teria sido observada para o grupo de tratamento, caso esse não tivesse sido beneficiado pela política pública analisada. Embora essa hipótese não possa ser testada, caso estejam disponíveis dados de outros períodos anteriores à intervenção, pode-se verificar se há indícios de que os dois grupos apresentavam tendências paralelas anteriormente;

2) Os grupos de tratamento e controle não devem ser afetados de forma distinta por outros fatores concomitantes à intervenção, e a composição dos mesmos não deve se alterar substancialmente ao longo do tempo, já que, caso contrário, as estimativas de impacto poderão ser viesadas.

Para ilustrar a intuição do método, sejam:

- GT_0: média do indicador de impacto para o grupo de tratamento no momento "antes";
- GT_1: média do indicador de impacto para o grupo de tratamento no momento "depois";
- GT_1': contrafactual hipotético correspondente à média do indicador de impacto para o grupo de tratamento no momento "depois", caso não tivesse sido tratado. Esse valor não pode ser observado na prática;
- GC_0: média do indicador de impacto para o grupo de comparação no momento "antes";
- GC_1: média do indicador de impacto para o grupo de comparação no momento "depois".

Para calcular o impacto usando diferença em diferenças, deve-se inicialmente calcular a diferença observada

[21] Baseado em Menezes Filho e Pinto (2017), Capítulo 4. Nesse capítulo, os autores apresentam maiores detalhes sobre esse método.

ao longo do tempo para cada um dos grupos, tal que:

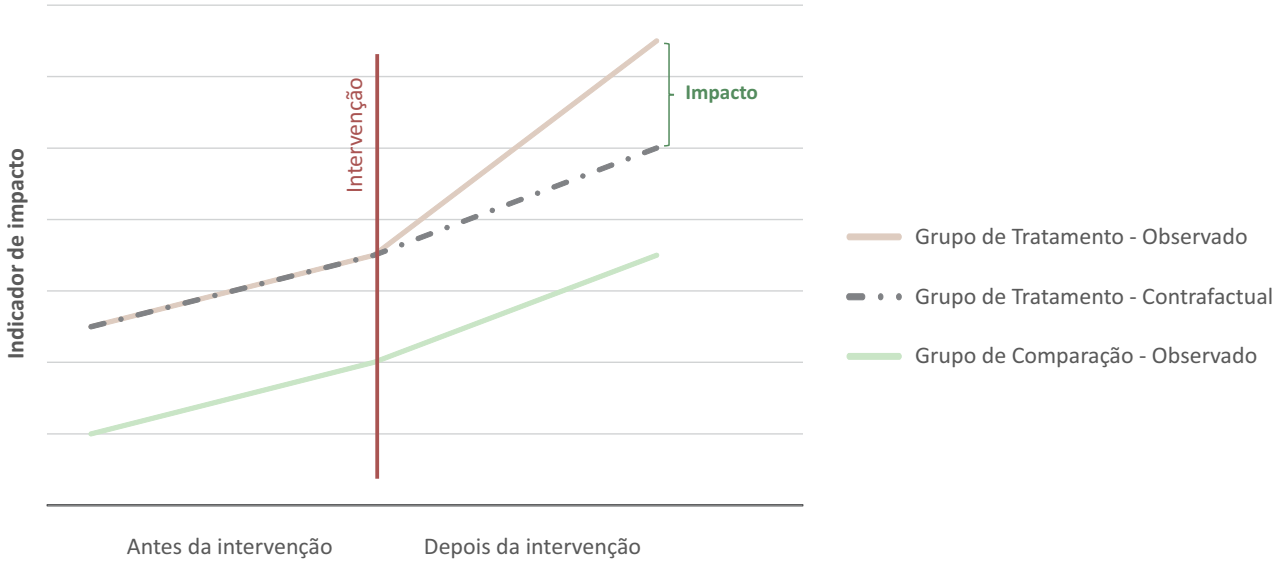
- Grupo de tratamento: $GT_1 - GT_0$
- Grupo de comparação: $GC_1 - GC_0$

Ao fazer isso, elimina-se o efeito de fatores fixos no tempo, sejam eles observáveis ou não observáveis, o que é uma das vantagens desse método. Depois, calcula-se o impacto a partir da diferença das diferenças anteriores:

$$Impacto = (GT_1 - GT_0) - (GC_1 - GC_0)$$

A intuição do método também está ilustrada na figura a seguir, que apresenta as trajetórias dos grupos de tratamento e de comparação ao longo do tempo (antes e depois da intervenção). A linha pontilhada representa a trajetória que se espera que o grupo de tratamento teria caso não tivesse sido beneficiado pela política pública em questão. Cabe notar que nesse método os grupos de tratamento e comparação não precisam necessariamente partir do mesmo nível no momento "antes", basta que eles apresentem trajetórias paralelas na ausência da intervenção.

Figura 1 | Ilustração do método de diferença em diferenças



Fonte: elaboração própria.

Boxe J

Estimação do impacto utilizando o método de diferença em diferenças

Para estimar o impacto de uma política pública utilizando-se o método de diferença em diferenças, pode-se utilizar a seguinte regressão[22]:

$$Y_{it} = \alpha + \beta Tratamento_i + \gamma Tempo_t + (\delta Tratamento_i \times Tempo_t) + \varepsilon_{it}$$

onde Y_{it} corresponde ao indicador de impacto para o indivíduo *i* no momento de tempo *t*, $Tratamento_i$ é uma variável binária indicativa de tratamento (assume o valor *1* se o indivíduo *i* é do grupo de tratamento e o valor *0* caso contrário), $Tempo_t$ é uma variável binária indicativa de tempo (assume o valor 1 para observações do momento "depois" e o valor *0* caso contrário) e ε_{it} é o termo de erro. O impacto será dado por δ.

Exemplo 3

Métodos estruturados de ensino

No artigo "*The Impact of Structured Teaching Methods on the Quality of Education in Brazil*", Leme et al. (2012) utilizam o método de diferença em diferenças para avaliar os impactos do uso de métodos estruturados de ensino sobre o desempenho e aprovação escolares dos alunos no Ensino Fundamental no Estado de São Paulo.

Intervenção avaliada

Conforme apresentado em Leme et al. (2012), o uso de métodos estruturados de ensino refere-se à contratação, por redes públicas de ensino, de instituições privadas provedoras de serviços educacionais para o oferecimento de uma proposta pedagógica, que geralmente inclui o desenvolvimento e provisão de material didático (ex: livros, exercícios para casa, gabaritos, materiais suplementares e planos de aulas para professores) e o treinamento de professores para a sua utilização. Esse tipo de intervenção tem como objetivo promover o aumento do desempenho

[22] Angrist e Pischke (2008) apresentam outras informações e detalhes sobre o método.

escolar por parte dos alunos. O uso de métodos estruturados de ensino se dá a partir de uma escolha da rede, sendo que os autores investigam os efeitos do mesmo para o caso de redes municipais de ensino.

Metodologia de avaliação de impacto

Para avaliar os impactos do uso de métodos estruturados sobre o desempenho e aprovação escolares dos alunos, Leme et al. (2012) utilizam dados longitudinais[23] de municípios no Estado de São Paulo, especificamente do Censo Escolar (INEP), Prova Brasil (INEP), dados coletados sobre se a rede municipal de ensino faz ou não uso de métodos estruturados, além de características sociodemográficas dos municípios obtidas a partir do Censo Demográfico (IBGE).

Os autores utilizam o método de diferença em diferenças e restringem a amostra para manter apenas municípios que não tinham adotado nenhum tipo de método estruturado de ensino até 2005 (393 municípios). O grupo de tratamento foi formado por municípios que adotaram métodos estruturados de ensino em 2006 ou 2007, totalizando 54 municípios para análises referentes à 4 série e 26 para análises referentes à 8 série. O grupo de controle foi formado pelos demais municípios da amostra (332 para análises referentes à 4 série e 100 para análises referentes à 8 série). O ano pré-intervenção considerado foi 2005 e o ano pós-intervenção para o qual os impactos são estimados foi 2007.

Principais resultados da avaliação de impacto

Alguns dos principais resultados reportados por Leme et al. (2012) sobre impactos do uso de métodos estruturados de ensino são::

- *Taxa de aprovação*: impacto positivo e estatisticamente significante (nível de 10%) apenas para a o segundo ciclo (5 à 8 séries) do Ensino Fundamental (3 pontos percentuais);
- *Notas na 4ª série*: impactos positivos e estatisticamente significantes (nível de 5%) sobre as notas de Matemática (5,3 pontos) e Língua Portuguesa (3,4 pontos);
- *Notas na 8ª série*: impacto positivo e estatisticamente significante (nível de 10%) apenas sobre as notas de Matemática (8,6 pontos).

[23] Dados longitudinais referem-se a informações que são observadas em diversos momentos de tempo para um mesmo conjunto de unidades de observação.

Variável instrumental

O método de avaliação de impacto utilizando variável instrumental pode ser utilizado quando a seleção para o tratamento da política pública em análise se baseia tanto em características observáveis quanto em não observáveis. Nesse método, utiliza-se uma fonte de variação exógena - *variável instrumental* - relacionada ao tratamento para identificar o impacto da política pública em análise. O desafio é identificar uma variável instrumental que cumpra com os requisitos expostos na sequência. Alguns exemplos de variáveis instrumentais frequentemente utilizadas são: (i) o status de tratamento definido originalmente a partir de um processo de aleatorização, em casos de experimentos em que nem todos os indivíduos seguem na prática o status que lhes foi atribuído; (ii) o oferecimento de encorajamento (ex: carta convite, divulgação extra) ou incentivos (ex: recompensa monetária) atribuído de forma aleatória; e (iii) características específicas do contexto (ex: índice de pluviosidade, tipo de colonização).

Considere, por exemplo, o caso de uma política pública de capacitação profissional em que a seleção tenha sido baseada em um processo de aleatorização dentre os candidatos inscritos, mas cuja participação/matrícula fosse voluntária por parte dos candidatos. É bastante provável que nem todos os indivíduos selecionados aleatoriamente para participar de fato participem, assim como também é possível que indivíduos selecionados para não participar acabem tendo acesso à intervenção. Pode-se dizer que a participação será definida na prática tanto pelo status de tratamento original quanto pela decisão individual de participar ou não (*autosseleção*).

Nesse caso, se for feita uma avaliação experimental considerando a definição original dos grupos de tratamento e comparação, o impacto estimado corresponderá ao chamado "efeito da intenção de tratar" que, pelos motivos expostos, será diferente do impacto da política sobre os participantes de fato. É possível, no

entanto, utilizar o status de tratamento original definido a partir do processo de aleatorização (Z_i) como variável instrumental para a participação no programa em questão (T_i) para recuperar o chamado "efeito médio local do tratamento" (*LATE – local average treatment effect*).

Ainda considerando esse contexto, pode haver quatro grupos de indivíduos[24]:

1) Indivíduos que participam sempre, mesmo que não tenham sido sorteados para participar (*always takers*);

2) Indivíduos que participam se tiverem sido sorteados para participar, mas não participam caso contrário (*compliers*);

3) Indivíduos que não participam se tiverem sido sorteados para participar, mas participam caso contrário (*defiers*);

4) Indivíduos que não participam nunca, mesmo que tenham sido sorteados para participar (*never takers*).

O efeito médio local do tratamento estimado a partir do uso de variáveis instrumentais diz respeito ao efeito específico para o grupo de *compliers*, uma vez que não é possível identificar o impacto para indivíduos de todos os quatro grupos.

Conforme exposto incialmente, a estimação de impactos de uma política pública a partir do método de variável instrumental requer justamente a existência de uma variável instrumental Z_i que seja correlacionada à participação na intervenção T_i e não seja correlacionada diretamente aos resultados de interesse, isto é, aos indicadores de impacto selecionados Y_i. Esse seria justamente o caso do exemplo anterior sobre o programa de capacitação profissional: o status original de tratamento definido a partir de um processo de aleatorização é correlacionado à participação no programa, mas não é diretamente correlacionado a um indicador de impacto de empregabilidade, por exemplo.

[24] Esses grupos são apresentadas e exemplificados em maiores detalhes em Menezes Filho e Pinto (2017), Capítulo 6.

Cabe ressaltar que o uso desse método é condicional à existência de uma variável instrumental que cumpra com os requisitos necessários e que os impactos estimados se referem especificamente ao grupo dos *compliers*, tal que usualmente não poderão ser generalizados para a população como um todo.

Boxe K

Hipóteses do método de variável instrumental

O método de variável instrumental requer que as seguintes hipóteses sejam válidas[25]:

1) *Alocação independente*: a variável instrumental não depende dos resultados ou tratamentos potenciais;

2) *Restrição de Exclusão*: a variável instrumental relaciona-se ao indicador de impacto apenas através da variação causada pela primeira na participação na intervenção.

3) *Monotonicidade*: a variável instrumental afeta todos os indivíduos no mesmo sentido. Essa hipótese implica que não há *defiers*.

Boxe L

Estimação do impacto utilizando o método de variável instrumental

Para estimar os impactos (LATE) via o método de variável instrumental[26], utiliza-se um procedimento de estimação de mínimos quadrados em dois estágios (MQ2E), tal que:

1) Estima-se $\hat{T}_i$ (participação na intervenção predita) a partir de:

$$T_i = \kappa + \delta Z_i + X_i'\theta + e_i$$

onde X_i' corresponde a um vetor de características observáveis relevantes, Z_i é a variável instrumental, T_i indica a participação na intervenção e e_i é o termo de erro.

[25] Essas hipóteses são apresentadas e discutidas em maiores detalhes em Menezes Filho e Pinto (2017), Capítulo 6.

[26] Angrist e Pischke (2008) apresentam outras informações e detalhes sobre o método.

2) Estima-se o impacto da política em questão (β_{LATE}) a partir de:

$$Y_i = \alpha + \beta_{LATE}\hat{T}_i + X_i'\Phi + \varepsilon_i$$

em que ε_i é o termo de erro, tendo as demais variáveis sido definidas previamente.

Exemplo 4

Eletrificação nas áreas rurais

No artigo "*Lighting and Homicides: Evaluating the Effect of an Electrification Policy in Rural Brazil on Violent Crime Reduction*", Arvate et al. (2017) utilizam o método de variável instrumental para avaliar os impactos da existência de iluminação sobre a ocorrência de homicídios em áreas rurais.

Intervenção e metodologia de avaliação de impacto

Arvate et al. (2017) investigam os efeitos da eletrificação, que possibilita a iluminação de vias públicas, sobre a taxa de homicídios em áreas rurais do Brasil. Os autores argumentam que existem vários mecanismos possíveis através dos quais a eletrificação pode afetar a ocorrência de crimes violentos. Por exemplo, é possível que a eletrificação leve a um aumento das atividades econômicas, aumentando a renda obtida através de atividades legais e reduzindo atividades ilegais e crimes. Outro argumento é que a eletrificação leva a um aumento da posse de bens domésticos, como televisão, que faz com que as pessoas passem menos tempo nas ruas, estando, assim, menos sujeitas a serem vítimas de crimes. O aumento da iluminação nas vias pode inibir a ocorrência de crimes pois a identificação do criminoso pode ser mais fácil na presença de luz, mas, por outro lado, também facilita a visualização de objetos de valor, o que pode levar a um aumento de atividades criminosas.

Para avaliar os impactos da iluminação sobre a taxa de homicídios (a cada 100.000 habitantes) em municípios de áreas rurais brasileiras, os autores utilizam uma abordagem de variável instrumental, uma vez que precisam de uma variação exógena para poder estimar os efeitos causais de interesse sem que haja viés. Assim, eles fazem uso dos critérios de elegibilidade e priorização para a expansão do programa Luz Para Todos, implementado pelo Governo

Federal ao longo da década de 2000 com o objetivo de prover iluminação para áreas rurais, para propor um instrumento para o acesso à iluminação (mensurado pelo percentual de domicílios com acesso à eletricidade no município). Os dados utilizados na avaliação são provenientes do DATASUS (Ministério da Saúde) e do Censo Demográfico (IBGE).

Principais resultados da avaliação de impacto

Arvate et al. (2017) encontram um efeito de redução na taxa de homicídios devido ao aumento da iluminação via eletrificação. Especificamente, em municípios da região Nordeste do país (a mais beneficiada pela expansão do programa Luz Para Todos), ao aumentar a cobertura de eletricidade de zero para cobertura completa, há uma redução de 92 casos de mortes violentas em áreas públicas e de 18 casos de mortes violentas em hospitais (a cada 100.000 habitantes).

Regressão descontínua

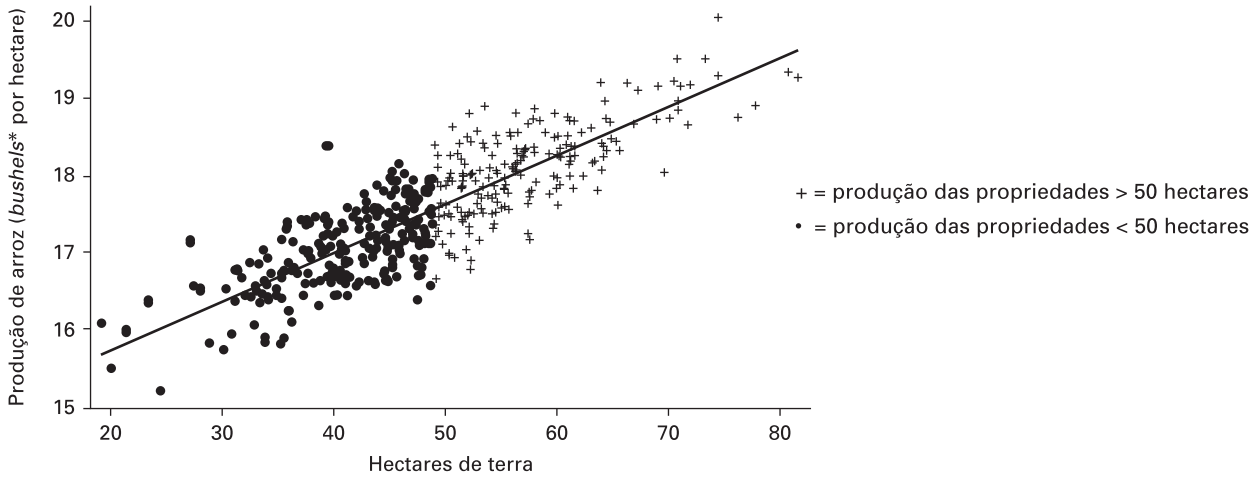
O método de regressão descontínua pode ser utilizado quando a seleção para participação na política pública em análise é baseada em um índice contínuo utilizado para classificar os indivíduos ou unidades (ex: renda, desempenho em exame de admissão, índice de vulnerabilidade) e existe um ponto de corte bem definido nesse índice, tal que todos os indivíduos ou unidades com desempenho acima/abaixo dele são considerados elegíveis/inelegíveis ao tratamento (Gertler et al, 2018).

Nesse método, a descontinuidade ao redor do ponto de corte do índice contínuo que define a elegibilidade ou participação é utilizada para que se possa estimar o impacto da política em análise naquela região específica, portanto, tratando-se também de um efeito médio local do tratamento (LATE). A intuição é que, por exemplo, indivíduos posicionados logo acima do ponto de corte, que foram considerados elegíveis e tiveram, portanto, acesso ao tratamento, são muito similares aos indivíduos que estavam posicionados logo abaixo desse ponto e que não foram, então, considerados elegíveis nem tratados. Assim, pode-se utilizar esses

indivíduos logo abaixo do ponto de corte para estimar o contrafactual dos indivíduos logo acima desse ponto, de forma que as diferenças observadas entre os grupos após a intervenção poderão ser atribuídas ao acesso à política em análise (*impacto*). Conforme discutido em Gertler et al. (2018), à medida que se aproximam do ponto de corte, as unidades logo acima ou logo abaixo do mesmo serão cada vez mais parecidas, tal que é quase como se o posicionamento delas ao redor desse ponto tivesse sido determinado de forma aleatória, remetendo a uma espécie de experimento local.

Os gráficos a seguir, reproduzidos de Gertler et al. (2018), ilustram o caso de uma aplicação de regressão descontínua considerando um programa agrícola cujo objetivo era promover o aumento da produção de arroz e para o qual eram elegíveis as propriedades com no máximo 50 hectares de terra. O primeiro gráfico apresenta a relação entre os hectares de terra e a produção de arroz das propriedades agrícolas em um momento anterior ao início do programa, onde nota-se que não há descontinuidades. Já o segundo gráfico apresenta a mesma relação, mas considerando um momento posterior ao programa, destacando como o impacto pode ser mensurado a partir da descontinuidade ("salto") observado ao redor do ponto de corte (50 hectares de terra).

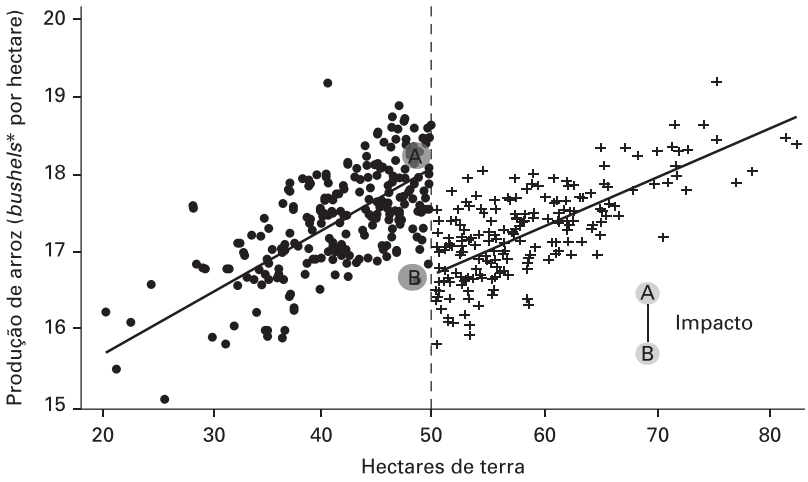
Figura 2 | Produção de arroz, propriedades menores versus propriedades maiores (linha de base)



Fonte: reproduzido de Gertler et al. (2018), página 128.

Figura 3 | Produção de arroz, propriedades menores versus propriedades maiores (período de acompanhamento)

+ = produção das propriedades > 50 hectares
• = produção das propriedades < 50 hectares



Fonte: reproduzido de Gertler et al. (2018), página 129.

Para estimar o impacto de uma política pública usando regressão descontínua, deve-se considerar qual é o caso pertinente: *sharp* ou *fuzzy*. No caso *sharp*, o status de tratamento T_i é definido de forma determinística a partir da variável contínua Z_i utilizada pela seleção para o tratamento (*running variable*), tal que $T_i = 1$ se $Z_i \leq c$ e $T_i = 0$ se $Z_i > c$, por exemplo. Nesse caso, a probabilidade de ser tratado salta de *0* para *1* no ponto de corte c.

Já no caso *fuzzy*, a probabilidade de ser tratado também é uma função de Z_i, mas não é definida de maneira determinística por ela:

$$P[T_i = 1|Z_i] = \begin{cases} g_0(Z_i) \text{ se } Z_i \leq c \\ g_1(Z_i) \text{ se } Z_i > c \end{cases} \quad \text{em que } g_0(Z_i) \neq g_1(Z_i)$$

Boxe M

Hipóteses do método de regressão descontínua

Conforme detalhado em Menezes Filho e Pinto (2017) e usando notação similar, o método de regressão descontínua baseia-se nas seguintes hipóteses:

1) *Continuidade*: $E[Y_i^1|Z_i=c]$ e $E[Y_i^0|Z_i=c]$ são contínuas em Z_i e ao redor do ponto de corte $Z_i = c$.

2) *Ignorabilidade local*: $(Y_i^1, Y_i^0) \perp T_i|Z_i = c$, tal que ao redor do ponto de corte $Z_i = c$ é como se o tratamento tivesse sido determinado de forma aleatória.

No caso *fuzzy*, há ainda uma hipótese adicional, a de *monotonicidade*, que expressa que o status potencial de tratamento em $Z = c$ é uma função não decrescente em Z.

Nos dois casos, o impacto é estimado a partir das diferenças observadas no indicador de impacto no momento posterior à intervenção entre os dois grupos (tratado e não tratado) na vizinhança do ponto de corte *c*. Esse processo envolve a definição do tamanho da janela ao redor do ponto de corte que será considerada, que dependerá da quantidade de unidades em cada lado e do balanceamento de características relevantes entre os grupos tratado e não tratado, o que envolverá um *trade-off* entre viés e variância. Também é preciso definir uma forma funcional para modelar a relação entre o indicador de impacto *Y* e a *running variable Z*. É importante verificar a sensibilidade dos resultados da avaliação à escolha do tamanho da janela e da forma funcional, conforme discutido em Gertler et al. (2018).

Boxe N

Estimação do impacto utilizando o método de regressão descontínua

Conforme detalhado em Menezes Filho e Pinto (2017) e usando notação similar, no caso *sharp* de regressão descontínua, o impacto pode ser estimado a partir de regressões lineares locais em cada lado do ponto de corte c, tal que:

$$Y_i = \alpha_{abaixo} + \beta_{abaixo}\,(Z_i - c) + e_i \;\; se \;\; c - h < Z < c$$

$$Y_i = \alpha_{acima} + \beta_{acima}\,(Z_i - c) + \epsilon_i \;\; se \;\; c \leq Z < c + h$$

em que e_i e ϵ_i correspondem a termos de erro e h refere-se ao tamanho da janela selecionada, tendo as demais variáveis sido definidas previamente. O impacto, que corresponde a um efeito médio local do tratamento (LATE) na vizinhança do ponto de corte *c* será dado pela diferença entre os interceptos nas regressões anteriores:

$$Efeito\ médio\ local\ do\ tratamento\ (LATE)_{sharp} = \alpha_{acima} - \alpha_{abaixo}$$

Já no caso *fuzzy* de regressão descontínua[27], utiliza-se uma abordagem similar à do caso anterior, mas também é necessário estimar a probabilidade de tratamento em cada lado do ponto de corte *c*, também utilizando regressões lineares locais:

$$T_i = \gamma_{abaixo} + \delta_{abaixo}\,(Z_i - c) + u_i \;\; se \;\; c - h < Z < c$$

$$T_i = \gamma_{acima} + \delta_{acima}\,(Z_i - c) + \mu_i \;\; se \;\; c \leq Z < c + h$$

onde u_i e μ_i correspondem a termos de erro, tendo as demais variáveis sido definidas previamente. O impacto (efeito médio local do tratamento - LATE) nesse caso será dado pela razão:

$$Efeito\ médio\ local\ do\ tratamento\ (LATE)_{fuzzy} = (\alpha_{acima} - \alpha_{abaixo}) / (\gamma_{acima} - \gamma_{abaixo})$$

em que α_{acima}, α_{abaixo}, γ_{acima} e γ_{abaixo} correspondem aos interceptos das regressões anteriores.

Conforme discutido em Gertler et al. (2018), é importante ressaltar que, assim como no caso do método de variável instrumental, os impactos estimados a partir do uso de regressão descontínua tratam-se de efeitos médios locais de tratamento, nesse caso, na vizinhança do ponto de corte que define

[27] Menezes Filho e Pinto (2017) detalham no Capítulo 7 o uso de regressão descontínua e as estimações envolvidas.

o acesso à política em questão. Dessa forma, os resultados não podem ser generalizados para indivíduos ou unidades distantes desse ponto de corte. Outro ponto de atenção para o uso desse método é a possibilidade de manipulação pelas unidades candidatas de seu desempenho no índice que define a elegibilidade/participação (*running variable*). O método de regressão descontínua somente deve ser utilizado quando não há possibilidade de manipulação; do contrário, proverá estimativas viesadas dos impactos da política pública avaliada.

Exemplo 5

Programa de Gestão Escolar por Resultados

No artigo " *The impact of school management practices on educational performance: Evidence from public schools in São Paulo*", Tavares (2015) utiliza o método de regressão descontínua para avaliar os impactos de um programa de gestão escolar sobre o desempenho dos alunos da 8 série no Estado de São Paulo.

Intervenção avaliada

Tavares (2015) investiga os impactos do Programa de Gestão Escolar por Resultados (PGER), que foi implementando no Estado de São Paulo e busca promover a adoção de práticas modernas de gestão nas escolas públicas estaduais. A autora explica que o programa inclui: treinamento administrativo para gestores escolares; diagnóstico, monitoramento e estabelecimento de metas para indicadores relacionados a aprendizagem; e desenvolvimento de planos de ação relacionados à gestão escolar. Em 2008, o programa foi implementado em caráter de programa-piloto e priorizou o atendimento das escolas com os piores resultados educacionais. A seleção das escolas atendidas foi baseada em uma regra arbitrária que estabelecia que todas as escolas na parte inferior da distribuição do resultado do IDESP de 2007 (especificamente, aquelas entre as 5% com menor desempenho) de cada nível de ensino (Ensino Fundamental I, Ensino Fundamental II e Ensino Médio)

receberiam o programa. O IDESP trata-se do Índice de Desenvolvimento da Educação do Estado de São Paulo, que pode assumir valores de 0 a 10 e é composto pela taxa média de aprovação e pela distribuição dos alunos em diferentes níveis de proficiência (abaixo do básico, básico, adequado e avançado).

Metodologia de avaliação de impacto

Para avaliar os impactos do PGER sobre o desempenho dos alunos, Tavares (2015) explora a variação exógena na probabilidade de participação no programa que é introduzida pela regra utilizada para selecionar as escolas a serem atendidas. Como a regra estabelecia que seriam tratadas as escolas que tivessem o IDESP entre os 5% mais baixos e essa análise era feita por nível de ensino, uma escola classificada como elegível devido ao IDESP associado à 8 série teria todas as suas séries tratadas, uma vez que o programa é implementado na escola como um todo, não sendo específico àquela série. Nesse contexto, a autora utilizada uma abordagem de regressão descontínua do tipo *fuzzy* adaptada[28], que possibilita identificar o efeito causal do programa utilizando a pontuação do IDESP como variável instrumental para a participação no mesmo. A autora consegue estimar qual seria o ganho em termos de desempenho escolar que os alunos em escolas que não receberam o PGER teriam tido caso suas escolas tivessem recebido o programa, com a ressalva de que o efeito estimado se trata de um efeito médio local específico para a vizinhança do ponto de corte estabelecido pela regra de seleção do programa. Foram utilizados dados do Censo Escolar (INEP) e do Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo SARESP (Secretaria da Educação do Estado de São Paulo).

Principais resultados da avaliação de impacto

Tavares (2015) mostra que, considerando o desempenho escolar de alunos na 8 série, o PGER apresenta, por disciplina:

- *Matemática*: impacto positivo e estatisticamente significante na proficiência (de 0,14 a 0,22 desvio padrão, o que equivale a aumentar o aprendizado anual de um aluno típico em 32% a 50%), sendo que esse resultado é proveniente exclusivamente do aumento da proficiência de alunos com maiores dificuldades acadêmicas (aqueles com nível de proficiência abaixo do básico);
- *Língua Portuguesa*: não foi encontrado impacto estatisticamente significante na proficiência.

[28] Tavares (2015) utilizada o método chamado de partially fuzzy RDD, que relaxa a hipótese de monotonicidade, conforme discutido e Battistin e Rettore (2008).

4.3. Tópicos em tipos de amostra, testes de hipóteses, efeito mínimo detectável e tamanho da amostra

4.3.1. Tipos de amostra

Avaliações de impacto de políticas públicas geralmente são realizadas a partir do estudo de uma *amostra estatística*, que corresponde a um subconjunto representativo de uma população de interesse, por exemplo, os indivíduos ou unidades elegíveis a um determinado programa. O uso de amostras para a realização de estudos em geral se dá devido a limitações do uso de informações sobre a população inteira, sejam elas fruto de restrições financeiras ou de tempo, uma vez que pode ser muito custoso ou inviável coletar as informações de todos os elementos de uma população, especialmente se o tamanho da mesma for muito grande. Dessa forma, para que estudos baseados no uso de uma amostra de fato informem sobre a população à qual se referem é imprescindível que a amostra utilizada seja representativa dessa população, especificamente, que preserve suas características médias. Alguns tipos de amostra frequentemente utilizados e que procuram garantir a representatividade da população são a *amostra aleatória simples* e a *amostra aleatória estratificada*.

No caso da *amostra aleatória simples*, observa-se o tamanho da população *(n)* e define-se o tamanho da amostra *(N)* a ser utilizada[29], tal que $N < n$. São selecionadas de forma aleatória N unidades da população ao, por exemplo, se atribuir um número aleatório para cada unidade, ordenar as unidades a partir desse número e selecionar as N primeiras. A principal vantagem desse tipo de amostra é que se N e n forem suficientemente grandes, garante-se que a amostra selecionada será representativa da população de interesse. No entanto, utilizar esse tipo de amostra pode não ser viável caso não se tenha uma lista com todas as unidades da população de interesse da qual se possa partir.

A *amostra aleatória estratificada* apresenta intuição e procedimentos parecidos aos do caso anterior. A diferença está no fato de que o primeiro passo nesse caso é dividir a população de interesse em estratos (subgru-

[29] A definição do tamanho da amostra dependerá de uma série de fatores relacionados, conforme discutido nas subseções seguintes.

pos) para os quais se deseja garantir representatividade também. Os estratos devem apresentar homogeneidade interna e heterogeneidade entre eles, como nos casos de sexo, faixa etária ou região geográfica. Uma vez definidos os estratos, seleciona-se uma amostra aleatória em cada um deles, seguindo um procedimento análogo ao descrito para o caso da amostra aleatória simples. Novamente, se o tamanho dos estratos e das amostras individuais de cada um forem suficientemente grandes, garante-se que as amostras selecionadas serão representativas de cada um deles e também da população de interesse. Cabe ressaltar que o uso desse tipo de amostra é vantajoso pois reduz o erro amostral, aumentando a precisão dos resultados. No entanto, o tamanho final da amostra considerando todos os estratos tende a ser maior, havendo dessa forma maiores custos associados à coleta de dados caso seja feita pesquisa de campo para esse fim.

4.3.2. Testes de hipóteses

Conforme discutido anteriormente, em avaliações de impacto procura-se estimar o tamanho dos impactos gerados pela política pública avaliada além de verificar se eles são estatisticamente significantes, isto é, se, dado um nível de confiança pré-estabelecido, é possível garantir que o impacto é diferente de 0 (zero).

Para verificar a significância estatística utiliza-se o procedimento estatístico de *teste de hipótese*, descrito em detalhes em Morettin e Bussab (2017). Em particular, em uma avaliação de impacto testa-se se o impacto estimado (correspondente à diferença das médias dos grupos de tratamento e de comparação) é diferente do valor 0 (zero), considerando-se um nível de significância de referência. Usualmente, a hipótese nula é de que o impacto é igual a zero e a hipótese alternativa é de o impacto é diferente de zero. Existem dois tipos de erros que podem ser cometidos nesse processo:

- *Erro tipo I*: rejeita-se a hipótese nula quando na verdade ela é verdadeira (falso positivo – diz-se que a política pública tem impacto quando ela na verdade não tem);

- *Erro tipo II*: não se rejeita a hipótese nula quando na verdade ela é falsa (falso negativo – diz-se que a política pública não tem impacto quando ela na verdade tem).

Na prática, é possível limitar a probabilidade de cometer erros do tipo I a partir da escolha do nível de significância do teste (α). Um valor usualmente utilizado é o de $\alpha = 5\%$, tal que o coeficiente de confiança é igual a 0,95. A intuição é que se forem construídas 100 amostras aleatórias de mesmo tamanho para uma dada população e calculados seus intervalos de confiança para a média, 95 conterão a média populacional.

Os erros do tipo II estão relacionados ao chamado *poder do teste*, que corresponde à probabilidade de rejeitar a hipótese nula quando ela é realmente falsa, dado um valor para a média populacional[30]. Um valor de referência usualmente utilizado para o poder do teste é de $1 - \kappa = 80\%$, em que κ é a probabilidade de incorrer em um erro do tipo II. Se em uma avaliação de impacto o poder do teste é pequeno, as conclusões obtidas a partir dela serão pouco informativas, uma vez que se pode concluir erroneamente que a intervenção não teve impacto quando na verdade ela pode ter tido, mas não havia poder suficiente para estimá-lo com precisão.

Boxe O

Testes de hipóteses

Formalmente, a hipótese nula (H_0) de que o impacto é igual a zero e a hipótese alternativa (H_1) de o impacto é diferente de zero podem ser representadas da seguinte forma, respectivamente:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

em que β é o impacto a ser estimado.

[30] Para informações mais detalhadas, veja Morettin e Bussab (2017).

Uma vez definidas H_0 e H_1, adota-se uma regra de decisão baseada em uma estatística de teste, tal que a hipótese nula será rejeitada quando:

$$\frac{|\hat{\beta}|}{\sigma} > c$$

em que $\hat{\beta}$ é o impacto estimado e σ é o desvio-padrão do estimador. O valor crítico c é definido a partir do nível de significância α escolhido:

$$Pr(|Z| > c) = \alpha$$

em que Z segue uma distribuição t de Student, considerando que a variância do estimador deverá ser estimada. Diremos que o impacto estimado é *estatisticamente significativo* ou *estatisticamente diferente de zero* quando a hipótese nula for rejeitada.

O poder do teste está relacionado ao chamado *efeito mínimo detectável*, que corresponde ao menor efeito possível no indicador de impacto que se consegue detectar na avaliação, e ao *tamanho da amostra*, sendo que há um *trade-off* entre esses dois elementos, conforme será discutido a seguir.

4.3.3. Efeito Mínimo detectável e tamanho da amostra

No contexto particular de avaliações experimentais, conforme apresentado em Duflo *et al*. (2007), o *efeito mínimo detectável* (EMD) é uma medida do quanto a variável dependente precisa ser afetada pelo tratamento para que sejamos capazes de detectar esse efeito, condicional à escolha de um nível de significância e um poder de teste. Dessa forma, valores menores de EMD são desejáveis, já que isso significa aumentar a capacidade de estimar os impactos de interesse.

Em linhas gerais, o efeito mínimo detectável é afetado pelo tamanho da amostra (amostras maiores levam a valores mais baixos de EMD) e pela proporção de unidades tratadas na amostra (proporções mais próximas de 50% levam a valores mais baixos de EMD), além da variância populacional envolvida e dos níveis de significância e de poder definidos. Dessa forma, é possível definir também o tamanho de amostra necessário para se atingir um determinado nível de EMD desejado.

Boxe P

Efeito mínimo detectável e tamanho da amostra

Formalmente, pode-se definir o *efeito mínimo detectável* (*EMD*) como:

$$EMD = (t_{1-k} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}}$$

em que t_{1-k} é o valor crítico da tabela da distribuição t correspondente a um poder de teste de $(1-\kappa)$; t_{α} é o valor crítico da tabela da distribuição t correspondente a um nível de significância estatística de α; P corresponde à proporção da amostra que recebeu tratamento; σ^2 corresponde à variância populacional; e N corresponde ao tamanho da amostra. Para o caso particular de $(1-\kappa) = 80\%$ e $\alpha = 5\%$, tem-se que $t_{0,8} = 0,84$ e $t_{0,05} = 1,96$.

Se os termos dessa equação forem rearranjados, é possível definir o tamanho da amostra (N) em função do efeito mínimo detectável (*EMD*), tal que:

$$N = \frac{(t_{1-k} + t_{\alpha})^2 \sigma^2}{P(1 - P) EMD^2}$$

Tanto no caso do cálculo do *EMD* quanto no de N: (i) o poder e o nível de confiança $t_{1-\kappa}$ e t_{α} são escolhidos pelo avaliador a partir de valores de referência usualmente utilizados; (ii) a proporção da amostra que recebeu o tratamento (P) é dada; e (iii) a variância populacional (σ^2) deve ser aproximada, utilizando, por exemplo, valores observados a partir do uso de estudo piloto com uma subamostra ou de informações proveniente de outras fontes. Para calcular o *EMD* toma-se N como dado e vice-versa. Observando as duas equações, nota-se que: (a) quanto maior for N, menor será o *EMD*; (b) quanto menor for o *EMD* requerido, maior deverá ser N; e (c) quando a amostra é igualmente dividida entre os grupos de tratamento e de comparação de forma que $P = 50\%$, a expressão $P(1-P)$ é maximizada, de forma que são minimizados tanto o *EMD* quanto N, mantendo-se os demais elementos constantes.

Para fazer o cálculo de poder na prática, é possível utilizar ferramentas disponíveis *online* que oferecem calculadoras de poder e simuladores do tamanho da amostra necessário em diferentes casos[31]. Por fim, Duflo *et al*. (2007) e Djimeu e Houndolo (2016) discutem o cálculo do efeito mínimo detectável em outros contextos de avaliação experimental, por exemplo, quando a aleatorização foi feita no nível de grupo ou quando a participação na intervenção está sujeita à escolha individual das unidades participantes do processo de aleatorização, tal que nem todas se comportam conforme o status pré-determinado de forma aleatória.

4.4. Tópicos em pesquisa de campo para avaliação de impacto

Caso a avaliação de impacto preveja o uso de dados de fontes primárias, isto é, dados coletados especificamente para uso na avaliação da política pública em questão, é preciso planejar e executar a coleta dos mesmos. Nesse contexto, é necessário desenhar um instrumento de coleta, planejar e implementar a pesquisa de campo e estabelecer diretrizes para a sistematização das informações coletadas[32].

Na presente subseção, serão discutidas algumas recomendações e boas práticas relacionadas a esses tópicos.

4.4.1. Instrumentos de coleta

O desenho dos instrumentos de coleta tem como primeiro passo a definição da forma como os dados serão coletados. Algumas formas comumente usadas são: (a) entrevista presencial; (b) entrevista por telefone; e (c) questionário *online*. Conforme discutido em Newcomer et al. (2015), a escolha da forma de coleta envolve uma série de fatores, dentre eles[33]:

- **Adequação do método ao público pesquisado**

Exemplo: um questionário *online* é inadequado para pesquisas com populações com pouco acesso à internet

[31] Um exemplo desse tipo de ferramenta está disponível em https://www.iadb.org/en/evaluationhub.

[32] Gertler et al. (2018, Capítulo 16) e Vermeersch et al. (2012, Módulo 4) discutem diversos aspectos relacionados à coleta de dados e à contratação de empresas especializadas para tal.

[33] Newcomer et al. (2015) apresentam no Capítulo 14 uma extensa discussão sobre esses e outros formatos de coleta de dados, bem como o trade-off envolvido considerando diversos fatores relevantes.

(ex: população em área rural, população vulnerável).

- **Adequação do método ao tipo de informação que será coletada**

Exemplo: caso seja preciso utilizar recursos visuais, entrevistas presenciais ou questionários *online* são preferíveis. Caso esteja prevista a inclusão de perguntas abertas ou complexas, que exigem respostas mais elaboradas, é preferível utilizar entrevistas presenciais. Por outro lado, caso sejam abordados tópicos sensíveis, entrevistas presenciais podem não ser a melhor opção, pois os respondentes podem ficar menos confortáveis para responder.

- **Adequação do método ao volume de informação que será coletada**

Exemplo: caso seja preciso coletar um volume de informações muito grande, questionários *online* e entrevistas por telefone não são recomendados, pois a probabilidade de o respondente desistir de participar da pesquisa aumenta muito.

- **Taxa de resposta esperada**

Exemplo: questionários *online* e entrevistas por telefone apresentam taxas médias de resposta muito menores que as de entrevistas presenciais.

- **Tempo disponível para a pesquisa de campo**

Exemplo: a duração da pesquisa de campo tende a ser muito maior nos casos de entrevistas presenciais ou por telefone do que quando são aplicados questionários *online*, justamente pela maior complexidade da logística envolvida (ex: transporte, necessidade de entrevistador).

- **Custos do campo**

Exemplo: entrevistas presenciais apresentam custos de coleta associados muito maiores que entrevistas por telefone ou questionários *online,* sendo que a última opção geralmente tem menor custo.

Uma vez definida a forma como os dados serão coletados, é preciso desenhar o instrumento de coleta

propriamente dito, o que envolve a definição dos conteúdos a serem abordados e as perguntas para tal. Ao definir as perguntas, é muito importante considerar o perfil dos respondentes e usar linguagem apropriada (ex: entrevista com jovens vis a vis entrevistas com adultos), bem como garantir que não serão utilizados termos desconhecidos pelos respondentes que possam prejudicar a qualidade das informações coletadas (ex: siglas, termos técnicos).

As perguntas devem ser breves, claras e objetivas, podendo ser facilmente compreendidas pelos participantes da pesquisa. Devem ser incluídas apenas perguntas cujas respostas trarão informações relevantes para a avaliação, do contrário, o questionário poderá ficar muito extenso, podendo comprometer a qualidade das informações coletadas já que aumentam as probabilidades de desistência e de provisão de informações incorretas ou imprecisas pelos respondentes.

O uso de perguntas fechadas que apresentem opções limitadas e pré-definidas de resposta (ex: múltipla escolha, escalas de grau de concordância) é preferível, pois facilita a sistematização e o uso dos dados no contexto de avaliações de impacto. Ressalta-se, no entanto, que no caso de perguntas fechadas é extremamente importante que as alternativas de resposta oferecidas contemplem todas as opções possíveis e que sejam mutuamente excludentes, exceto quando se deseja que o respondente selecione todos os itens de uma lista que sejam pertinentes.

Ainda em relação à formulação das perguntas, deve-se evitar o uso de perguntas extensas e/ou compostas (ex: "Qual é seu grau de satisfação com o curso e com a infraestrutura da sala de aula?") e também de perguntas que possam induzir a uma resposta específica (ex: "Depois de receber o subsídio para a compra de insumos, a produtividade do seu empreendimento aumentou?"). O uso de perguntas retroativas deve ser feito com cautela, pois estarão sujeitas à memória dos respondentes, o que pode comprometer a precisão das informa-

ções a depender do que for perguntado. Por exemplo, é provável que uma pessoa se lembre se estava trabalhando há 5 anos, mas possivelmente ela não se lembrará com exatidão qual era seu salário. Caso sejam utilizadas perguntas retroativas, é recomendado que o intervalo de tempo envolvido não seja muito longo e que se procure vincular as perguntas a algum evento relevante que ajude os respondentes a se situarem na época à qual a pergunta se refere (ex: eleições, inauguração de uma obra pública de grande porte). Newcomer et al. (2015) apresentam no capítulo 14 outras considerações relevantes para a formulação de perguntas adequadas para instrumentos de coleta de dados.

4.4.2. Pesquisa de campo

Para que a pesquisa de campo seja bem-sucedida, é fundamental que seja planejada em detalhes e considerando todas as etapas envolvidas, como:

(i) Elaboração dos instrumentos a serem utilizados;

(ii) Elaboração da documentação necessária (ex: termo de sigilo, formulário de consentimento, documentos de apresentação da pesquisa e de identificação dos entrevistadores, caso se aplique) [34];

(iii) Treinamento dos entrevistadores (caso se aplique);

(iv) Testes dos instrumentos;

(v) Logística de campo (ex: material para a pesquisa, impressão de questionários, configuração de *tablets*, planejamento de viagens, agendamento de entrevistas);

(vi) Controle de qualidade (ex: repetição de entrevistas, conferência da consistência dos dados coletados); e

(vii) Sistematização das informações coletadas.

Embora o planejamento para uma pesquisa de campo seja específico a seu contexto, há algumas recomendações úteis para todos os casos. Independentemente da forma de coleta de dados escolhida, é

[34] Vermeersch et al. (2012, Módulo 4) discutem aspectos relevantes para a elaboração de alguns desses documentos (ex: termo de confidencialidade, formulário de consentimento).

imprescindível que os instrumentos passem por pelo menos uma rodada de teste antes do início da pesquisa de campo (*pré-teste*). Deve ser selecionada uma subamostra representativa da população que será pesquisada para verificar o funcionamento dos procedimentos a serem adotados no campo na prática e se as perguntas estão adequadas, tanto do ponto de vista de compreensão quanto de conteúdo. Em particular, esse tipo de teste pode ser utilizado para checar se há problemas quanto aos termos e conceitos utilizados, se as opções de resposta estão adequadas e são suficientes, se o encadeamento lógico das perguntas está apropriado e ainda para mensurar o tempo necessário para responder ao questionário - informação que também é essencial para o planejamento das atividades. A partir das informações coletadas no teste, poderão ser feitos ajustes nos instrumentos e procedimentos de coleta. A depender dos resultados, pode ser realizada mais de uma rodada de testes.

Outro aspecto importante que independe do formato escolhido para a coleta de dados é a apresentação da pesquisa aos respondentes. Deve ficar claro para eles qual instituição está realizando a coleta de dados e para qual finalidade as informações serão utilizadas, deixando claro que tipo de sigilo será garantido. Essa apresentação é fundamental para incentivar os respondentes a participar da pesquisa e deixá-los confortáveis para tal.

Nos casos de entrevistas por telefone ou presenciais, os entrevistadores têm papel fundamental na qualidade dos dados coletados. Dentre a documentação necessária a ser providenciada, deve constar um documento de diretrizes ou manual para orientar o trabalho dos entrevistadores. Esse material deve ser detalhado e conter informações sobre todos os procedimentos a serem realizados durante a entrevista, por exemplo, a forma de apresentar a pesquisa, como fazer cada uma das perguntas, como lidar com situações adversas (ex: dúvidas, recusa) e como responder a perguntas frequentes. Além de receber esse material, os entrevistadores também precisam ser treinados antes do início do

campo, recomendando-se inclusive realizar rodadas de teste para as entrevistas.

Em qualquer formato de coleta de dados, devem estar previstos procedimentos de controle de qualidade. Conforme discutido no Capítulo 16 de Gertler et al. (2018), durante a realização da pesquisa de campo o controle de qualidade pode incluir a condução de tentativas adicionais de entrevista em casos de não resposta, acompanhamento de entrevistas por um supervisor e verificações com subamostras selecionadas aleatoriamente (ex: um supervisor refaz uma entrevista já conduzida para checar a consistência dos dados obtidos).

4.4.3. Sistematização das informações coletadas

Uma vez coletadas as informações a partir da pesquisa de campo, é necessário sistematizá-las na forma de um banco de dados para que sejam utilizadas posteriormente na avaliação de impacto. No caso de questionários *online* e de entrevistas realizadas com base em registros em dispositivos móveis, é comum que eles gerem um banco de dados automaticamente. Já nos casos em que são utilizados instrumentos em papel, será preciso digitar as respostas coletadas.

Nesse contexto, é preciso definir de antemão qual será o sistema a ser utilizado para o armazenamento das informações coletadas, bem como o destino dos registros originais no caso de instrumentos em papel, sendo recomendado que eles sejam preservados após a pesquisa de campo.

A sistematização das informações coletadas também envolve[35]: (i) a condução de rodadas de controle de qualidade, sobretudo quando são utilizados instrumentos em papel, já que pode haver erros durante o processo de digitação das respostas; (ii) a verificação da consistência dos dados (ex: eliminar informações duplicadas, verificar consistência entre respostas a itens diferentes); e (iii) o preparo da base de dados (ex: categorização de informações, limpeza dos dados) e de sua documentação para uso na avaliação de impacto.

[35] Gertler et al. (2018, Capítulo 16) e Newcomer et al. (2015, Capítulo 14) discutem outros aspectos relacionados à sistematização das informações coletadas em pesquisas de campo.

5 Avaliação de custo-benefício e custo-efetividade

Ao avaliar uma política pública, além de identificar os produtos, resultados e impactos promovidos por ela, é importante considerar os custos incorridos, tal que se possa verificar se os investimentos realizados são compensados pelos benefícios obtidos, por exemplo, ou qual foi o investimento necessário para alcançar determinados objetivos. Alguns exemplos são: investimento necessário por aluno para a obtenção de certificado em um curso profissionalizante, investimento necessário para evitar uma internação por febre amarela.

Essa é justamente a proposta das chamadas *avaliações de custo-benefício e de custo-efetividade*, que tem como objetivo confrontar os impactos gerados a partir de uma intervenção aos custos a ela associados. Nos dois casos, são utilizados os resultados obtidos em uma etapa anterior de avaliação de impactos e o os resultados obtidos são úteis para avaliar a eficácia das políticas e eventualmente comparar políticas alternativas que tenham objetivos similares.

5.1. Avaliação de custo-benefício

5.1.1. Descrição e indicadores utilizados

Avaliações de custo-benefício baseiam-se na comparação entre: (i) a totalidade de *custos* incorridos para viabilizar a política pública em análise; e (ii) os *benefícios* gerados pela política, que correspondem aos impactos por ela produzidos. Tanto os custos quanto os benefícios devem ser mensurados em *termos monetários*, isto é, em Reais. No caso dos custos, é bastante comum que todos ou pelo menos a maioria deles estejam originalmente definidos em termos monetários. Já no caso dos benefícios, esse não necessariamente é o caso, havendo a necessidade de monetizá-los quando assim for possível.

Conforme discutido em Newcomer et al. (2015), a avaliação de custo-benefício é recomendada quando se deseja avaliar apenas uma intervenção, para verificar se os benefícios gerados por ela excedem os custos incorridos para sua implementação, ou quando se deseja

[11] Avaliação de impacto é um tipo de avaliação ***ex post*** que busca identificar e mensurar os efeitos que podem ser atribuídos exclusivamente à política. Esse tópico é abordado no volume "E quando a política está em andamento? Avaliação ***ex post***!".

comparar diferentes políticas alternativas para identificar qual delas traz maiores benefícios para a sociedade a menores custos. A principal limitação desse tipo de avaliação relaciona-se à necessidade de medir tanto os custos quanto os benefícios em termos monetários, o que muitas vezes dependerá da adoção de hipóteses que podem estar sujeitas a críticas.

A comparação de custos e benefícios em termos monetários na avaliação de custo-benefício pode ser feita utilizando-se indicadores como[36]:

- ***Benefício Líquido (BL)***, tal que:

$$BL = Benefício\ Total - Custo\ Total$$

Nesse caso, o *BL* indica qual é o benefício gerado pela intervenção líquido dos custos associados a ela, em Reais. Se $BL > 0$, isso indica que a intervenção é recomendável, pois seus benefícios excedem seus custos.

- ***Razão Custo-Benefício (RCB)***[37], tal que:

$$RCE = \frac{Custo\ Total}{Benefício\ Total}$$

Nesse caso, a *RCB* informa qual é o investimento necessário por unidade de benefício gerado, em Reais. Se $RCB < 1$, isso indica que a intervenção é recomendável, pois seus benefícios excedem seus custos.

5.1.2. Levantamento de custos e de benefícios

Para que a avaliação de custo-benefício seja de fato informativa e possa contribuir para a tomada de decisões, devem ser levantados todos os custos e benefícios associados à política pública analisada. Muitas vezes alguns itens são difíceis de serem mensurados ou monetizados, mas devem ser elencados para que a análise seja completa, ainda que aqueles que sejam negligenciáveis possam ser desconsiderados no cálculo e apenas mencionados na discussão dos resultados (Newcomer et al., 2015, Capítulo 24). A seguir, será discutido o levantamento de custos e de benefícios e suas respectivas monetizações.

[36] Menezes Filho e Pinto (2017) apresentam no Capítulo 8 os indicadores a seguir (benefício líquido e razão custo-benefício) e outras alternativas de maneira detalhada.

[37] Em avaliações de custo-benefício, também é comum utilizar o indicador de Razão Benefício-Custo (RBC), em que se divide o Benefício Total pelo Custo Total (RBC=1/RCB). Nesse caso, o indicador mede qual foi o valor do benefício gerado (em Reais) para cada Real investido. Analogamente, a intervenção será recomendável se RBC > 1.

Custos

No contexto da avaliação de custo-benefício, devem ser considerados todos os custos associados à execução da política, sendo eles diretos ou indiretos[38]. Devem ser incluídos, por exemplo, custos associados à aquisição ou investimento em bens de capital, à aquisição de materiais de consumo, pagamento de serviços básicos (ex: água, luz), remuneração da equipe envolvida, entre outros. No caso de custos administrativos, caso a política não tenha uma equipe própria que realize essas funções, pode ser feito um rateio dos custos associados à equipe administrativa do órgão responsável pela política, considerando a parcela de tempo desses indivíduos que fica alocada para o trabalho na administração da política em análise. Em geral, esses tipos de custo são originalmente mensurados em termos monetários.

No levantamento de custos também devem ser incluídos os chamados *custos de oportunidade*. Esses custos se referem ao valor associado a insumos pelos quais não necessariamente é preciso pagar pelo uso (não há transação financeira envolvida), mas que, caso não fossem empregados na intervenção em análise, poderiam ter usos alternativos. Alguns exemplos são o tempo de trabalho oferecido por voluntários, o tempo utilizado pelos beneficiários para participar do programa e o uso de recursos próprios como prédios e veículos. Nesses casos, será preciso atribuir um valor a esses custos, o que geralmente é feito utilizando valores de mercado de referência (ex: valor do salário médio que precisaria ser pago para que alguém executasse as tarefas realizadas pelos voluntários, valor proporcional do salário médio que os participantes deixam de ganhar por dedicarem seu tempo à participação na intervenção, valor médio do aluguel de espaços semelhantes aos necessários para a realização da intervenção que teria de ser pago caso não fosse utilizado um espaço próprio/cedido).

[38] J-PAL (2016) sintetizam aspectos relevantes sobre o processo de levantamento de custos no contexto de avaliações de intervenções.

Benefícios

Em uma avaliação de custo-benefício, os benefícios dizem respeito aos impactos gerados pela política pública em questão mensurados em termos monetários (Reais). Em muitos casos, a unidade utilizada para medir os impactos não estará em termos monetários, de forma que será preciso propor uma metodologia para converter esses valores, o que envolverá a adoção de hipóteses e o uso de valores de referência. Esse processo é comumente chamado de *monetização*.

Newcomer et al. (2015, Capítulo 24) discutem algumas técnicas utilizadas para monetizar os impactos de uma intervenção, destacando-se:

- *Aumento da produtividade*: são utilizadas informações de valores de salários e lucros de referência.
- *Disposição a pagar*: são utilizadas informações do valor a ser pago pelos serviços prestados pela política em análise em um mercado privado similar ou mesmo questionando os beneficiários que a utilizam quanto eles pagariam por eles, embora haja risco de viés nesse último caso.
- *Custos evitados*: são utilizadas informações de dados históricos e de valores de referência de mercado para se calcular quanto se conseguiu poupar em termos monetários graças aos impactos promovidos pela política em questão. Por exemplo, um impacto de redução da infecção por febre amarela terá associado a ele uma série de gastos evitados com medicamentos, internações e perda de produtividade, dentre outros.

Para contabilizar os benefícios totais, devem ser levados em consideração o tamanho da amostra ou o número de indivíduos/unidades beneficiados pela política[39] e a duração dos benefícios[40].

[39] A depender dos parâmetros estimados na avaliação de impacto realizada anteriormente.
[40] Dhaliwal et al. (2013) discutem esses aspectos em detalhes.

5.1.3. Cálculo de custos e benefícios totais

Uma vez identificados todos os custos e benefícios a serem levados em consideração na avaliação de custo-benefício, para calcular seus valores totais é preciso ainda:

i) Identificar os valores e o momento no tempo em que cada item é observado:

Para os custos, é preciso identificar para cada insumo utilizado quando seu uso ocorre e quais são os valores associados, sendo comum que eles estejam concentrados no início do intervalo de tempo considerado na avaliação. Já para os benefícios, além de identificar os valores e momentos do tempo em que ocorrem, também é preciso considerar qual é a duração esperada dos impactos e como suas magnitudes se comportam ao longo do tempo (ex: se dissipam, permanecem estáveis, aumentam). Recomenda-se o uso de ferramentas como o *fluxo de caix*a para facilitar a visualização e os cálculos posteriores[41].

O processo de monetização e de definição das durações no tempo de cada item envolve hipóteses que devem ser fundamentadas e explícitas de forma detalhada na descrição da metodologia da avaliação de custo-benefício, para que possam ser analisadas pelo gestor que usará os resultados da avaliação. Além disso, é recomendado que sejam feitas análises de sensibilidade dos resultados a essas hipóteses, conforme discutido posteriormente.

ii) Calcular os valores presentes do custo total e do benefício total:

Todos os valores de custos e de benefícios devem ser trazidos a valor presente, isto é, todos devem se referir a um mesmo momento de tempo. É comum, por exemplo, utilizar o ano de início da intervenção como referência (uma alternativa seria utilizar o ano de início da avaliação). Isso é necessário para que se possa agregar todos os custos *(Custo Total)* e todos os benefícios *(Benefício Total)* e assim compará-los utilizando algum dos indicadores apresentados anteriormente.

[41] Menezes Filho e Pinto (2017, Capítulo 8) discutem detalhadamente o uso de ferramentas de fluxo de caixa e exemplos de monetização.

Boxe Q

Cálculo do valor presente

Para calcular o valor presente dos custos e dos benefícios, utilizam-se as seguintes fórmulas, a depender do período de referência adotado:

- Custos e benefícios posteriores (futuros) ao período de referência:

$$\textit{Valor Presente do Custo Total Futuro} = \sum_{t=0}^{T} \frac{\text{Custo}_t}{(1+s)^t}$$

$$\textit{Valor Presente do Benefício Total Futuro} = \sum_{t=0}^{T} \frac{\text{Benefício}_t}{(1+s)^t}$$

- Custos e benefícios anteriores (passados) ao período de referência:

$$\textit{Valor Presente do Custo Total Passado} = \sum_{t=0}^{T} \text{Custo}_t (1+s)^t$$

$$\textit{Valor Presente do Benefício Total Passado} = \sum_{t=0}^{T} \text{Benefício}_t (1+s)^t$$

Em que t é o momento de tempo; T é o valor máximo que t pode assumir sendo, portanto, o horizonte temporal total considerado na avaliação[42]; $Custo_t$ é o total de custos ocorridos em t; $Benefício_t$ é o total de benefícios ocorridos em t; e s é a chamada taxa social de desconto, que desempenha um papel análogo ao de uma taxa de juros típica e cuja escolha deverá ser analisada caso a caso[43].

Newcomer et al. (2015) apresentam no Capítulo 24 um exemplo de levantamento de custos e de benefícios para um programa de prevenção à evasão escolar no ensino médio. Também é apresentado um quadro que exemplifica a descrição das hipóteses envolvidas no cálculo dos custos desse programa e os cálculos de custos e benefícios totais.

[42] Newcomer et al. (2015, Capítulo 24) recomendam considerar um horizonte temporal que seja suficiente para captar a maioria dos custos e benefícios da intervenção em análise.

[43] Newcomer et al. (2015, Capítulo 24) e Menezes Filho e Pinto (2017, Capítulo 8) discutem aspectos relacionados à escolha da taxa social de desconto para a avaliação.

5.1.4. Análise de sensibilidade

Recomenda-se que a avaliação de custo-benefício inclua análises de sensibilidade dos resultados aos parâmetros envolvidos, dentre eles:

- Estimativas de impacto (ex: considerar o intervalo de confiança estimado para os impactos)[44];
- Hipóteses utilizadas, tanto para os custos quanto para os benefícios:
  - Valores de referência utilizados nos casos de monetização;
  - Duração e comportamento da magnitude dos benefícios ao longo do tempo;
  - Taxa social de desconto utilizada.

Em suma, a análise de sensibilidade envolve a criação de cenários e pode tomar duas formas distintas, conforme apresentado em Newcomer et al. (2015, Capítulo 24):

a) Análise de sensibilidade parcial: altera-se um parâmetro por vez, mantendo todos os demais inalterados;

b) Análise de sensibilidade extrema: alteram-se diversos parâmetros simultaneamente, procurando identificar cenários do tipo "melhor caso possível" e "pior caso possível".

Esses dois tipos de análises produzem informações de naturezas distintas, sendo essas complementares. A análise considerada mais informativa será específica para cada caso.

[44] Dhaliwal et al. (2013) discutem detalhadamente a questão de como considerar imprecisões das estimativas de impacto em avalições de custo-benefício e de custo-efetividade.

Exemplo 1

Programas para a juventude

No estudo "*Benefits and costs of prevention and early intervention programs for youth*"[45] realizado pelo *Washington State Institute for Public Policy*[46] (*WSIPP*), são analisados os benefícios e os custos de diversos programas endereçados ao público jovem e que tenham como objetivo: (i) reduzir a incidência de crimes; (ii) reduzir a incidência de uso abusivo de substâncias; (iii) melhorar resultados educacionais (ex: aumentar desempenho em testes padronizados e índices de aprovação); (iv) reduzir a incidência de casos de gravidez durante a adolescência; (v) reduzir a incidência de tentativas de suicídio entre adolescentes; (vi) reduzir a incidência de abuso ou negligência infantil; ou (vii) reduzir a incidência de violência doméstica.[47]

Um dos programas analisados nesse estudo foi o *Home Instruction Program for Preschool Youngsters (HIPPY)*[48], uma intervenção que tem como objetivo promover o desenvolvimento infantil por meio da provisão de instruções para os pais sobre como estimular e ensinar seus filhos durante os primeiros anos de vida. O programa é voltado para famílias em que haja crianças de 3 anos de idade cujos pais apresentem baixa escolaridade. Nele, instrutores fazem visitas domiciliares quinzenais às famílias beneficiadas, nas quais entregam livros e brinquedos educativos e oferecem instruções para os pais sobre como utilizá-los e como estimular o aprendizado infantil.

Aos et al. (2004) basearam-se em informações disponíveis no website do programa HIPPY para estimar os custos totais por beneficiário do mesmo, considerando o tempo médio de duração da intervenção (1,5 ano). Como custos do programa, foram incluídos os valores dos materiais e dos treinamentos realizados. O custo total estimado por beneficiário foi de US$ 1.837, sendo que os autores utilizaram o *Implicit Price Deflator (IPD) for Personal Consumption Expenditures* dos Estados Unidos como deflator para calcular os valores em dólares de 2003.

No caso dos benefícios, Aos et al. (2004) consideraram os impactos estimados por outros estudos referentes aos efeitos do HIPPY sobre notas em testes padronizados e calcularam que o efeito médio foi de 0,052 desvio-padrão[49]. Como esse impacto não foi medido em unidades monetárias, os autores adotaram algumas hipóteses e procedimentos para poder monetizá-lo:

[45] Aos et al. (2004).

[46] Mais informações, consulte: <http://www.wsipp.wa.gov>

[47] Aos et al. (2004) apresentam os resultados específicos para cada um dos diferentes programas analisados e também uma comparação entre esses resultados estimados.

[48] Todas as informações sobre a descrição do HIPPY apresentadas nesse exemplo foram baseadas em Aos et al. (2004).

[49] Aos et al. (2004) detalham a metodologia utilizada para calcular esse efeito médio com base nos efeitos estimados por outros artigos.

- Foi calculada a média de salários por idade a partir de dados de uma pesquisa suplementar ao censo demográfico dos Estados Unidos (*March 2002 Supplement to the Current Population Survey*) e foi calculado o fluxo de ganhos esperados em temos de salários;
- Foram utilizadas as estimações de outros estudos que calculam qual é a taxa média de retorno em termos de salário para um aumento de 1 desvio-padrão na nota no contexto americano (valor de referência usado: aumento de 12% no salário). Esse valor foi multiplicado pelo fluxo de ganhos esperados em temos de salários calculado no passo anterior, ajustando pelo impacto estimado;
- Foi calculado o valor presente do fluxo de ganhos esperados em temos de salários para a idade de 18 anos, considerando: (i) uma taxa de desconto de 3%; (ii) uma taxa de crescimento real anual dos salários de 0,5%; e (iii) que o período de duração dos benefícios corresponderia ao horizonte de 18 a 65 anos de idade.

O benefício total estimado[50] por beneficiário foi de $3.313 em dólares de 2003 (novamente, os autores utilizaram o *Implicit Price Deflator (IPD) for Personal Consumption Expenditures* dos Estados Unidos como deflator).

Assim, Aos et al. (2004) calculam que o *Benefício Líquido (BL)* do HIPPY foi de:

$$BL = \text{Benefício Total} - \text{Custo Total} = US\$\ 3.313 - US\$\ 1.837 = US\$\ 1.476$$

Por sua vez, a *Razão Benefício-Custo (RBC)* do HIPPY foi de:

$$RBC = \frac{US\$\ 3.313}{US\$\ 1.837} = 1,80$$

Como $BL > 0$ e $RBC > 1$, pode-se dizer que a intervenção é recomendável, pois seus benefícios excedem seus custos.

[50] Aos et al. (2004) detalham a metodologia utilizada para calcular o benefício total estimado.

5.2. Avaliação de custo-efetividade

Conforme discutido em Newcomer et al. (2015, Capítulo 24), a avaliação de custo-efetividade é recomendada quando se deseja comparar diferentes políticas alternativas que apresentem impactos nas mesmas dimensões, tal que se consegue verificar qual delas é capaz de gerar o maior impacto ao menor custo. Na prática, esse tipo de avaliação é muito utilizado quando a monetização dos impactos não é possível ou não é recomendada, pois é dependente de hipóteses questionáveis e pouco críveis, por exemplo.

Define-se a *Razão Custo-Efetividade (RCE)* como:

$$RCE = \frac{Custo\ Total}{Impacto}$$

A *RCE* mede o investimento necessário para obter cada unidade de impacto não monetário (ex: aluno que concluiu o ensino médio, vida salva). Nota-se que quanto maior for a *RCE*, menos eficiente é a intervenção.

Para calcular o *Custo Total* para a *RCE*, utilizam-se os mesmos passos e métodos descritos para o cálculo desse mesmo item no contexto da avaliação de custo-benefício. Já para calcular o *Impacto*, utilizam-se os resultados de uma avaliação de impacto anterior, também se devendo considerar o tamanho da amostra ou número de indivíduos/unidades beneficiados pela política e a duração dos benefícios ao longo do tempo.

Do ponto de vista metodológico, as principais diferenças entre a avaliação de custo-efetividade e a de custo-benefício é que na primeira os impactos não são monetizados e, caso haja impactos em diferentes dimensões, esses não são agregados em uma única medida. Nesse sentido, destaca-se que uma limitação associada à *RCE* é que ela foca em apenas uma medida de impacto por vez, o que pode não ser ideal quando se está avaliando programas com vários impactos em dimensões diferentes. Dhaliwal et al. (2013) discutem esse aspecto e exemplificam alguns casos em que é possível fazer um rateio dos custos entre os diferentes impactos promovidos pela intervenção, ressaltando que isso nem sempre é factível.

Por fim, assim como no caso da avaliação de custo-benefício, recomenda-se que sejam feitas análises de sensibilidade dos resultados da avaliação de custo-efetividade, de maneira análoga à avaliação de custo-benefício, descrita anteriormente.

Exemplo 2

Programa Balsakhi

O Programa Balsakhi[51] foi implementado em escolas municipais da Índia a partir de 1994 e foi avaliado no estudo de Banerjee et al. (2007). Trata-se de um programa de reforço escolar para crianças matriculadas na 3 e 4 séries que não atingiram o nível de proficiência esperado para essas séries. No programa, as crianças com essa característica passam cerca de metade do tempo de sua jornada diária na escola fora da sala de aula de sua turma regular para poder participar de atividades em outra turma específica. Nessa outra turma, são desenvolvidas atividades de reforço escolar com base em um currículo padronizado, proposto pelo próprio programa. Essas atividades são conduzidas por um professor – o *balsakhi* – , que é geralmente uma mulher jovem da própria comunidade, que recebe treinamento para tal.

Banerjee et al. (2007) avaliam os impactos do Programa Balsakhi utilizando o método experimental[52]. Para isso, os autores consideram a expansão do programa nos municípios de Vadodara e Mumbai, sendo que, nos dois casos, esta não ocorreu simultaneamente em todas as escolas, tendo a ordem de implementação sido definida a partir de um processo de aleatorização. Eles reportam que, no primeiro ano de implementação, o programa teve impacto de 0,138 desvio-padrão sobre a pontuação dos alunos em um teste padronizado que incluía conteúdos de matemática e linguagem.

A partir dos resultados obtidos na avaliação de impacto de Banerjee et al. (2007), o Abdul Latif Jameel Poverty Action Lab (J-PAL)[53] conduziu uma análise de custo-efetividade do Programa Balsakhi[54]. Para tal, foi necessário levantar de forma detalhada os custos do programa, que incluíram a remuneração de pessoal (salários dos gestores do programa, instrutores e *balsakhis*) e os gastos com transporte e materiais didáticos. Os valores dos itens de custos foram calculados inicialmente em moeda local (Rúpia Indiana) e depois foram convertidos para dólares

[51] Balsakhi significa "amigo da criança". Todas as informações sobre a descrição do Programa Balsakhi apresentadas nesse exemplo foram baseadas em Banerjee et al. (2007).

[52] Para mais informações sobre o método experimental de avaliação de impacto, ver o Capítulo 4 deste volume do Guia.

[53] Mais informações, consulte: <https://www.povertyactionlab.org/about-j-pal>.

[54] A análise completa do Programa Balsakhi, bem como de outros programas voltados a aumentar o aprendizado dos estudantes estão disponíveis em: <https://www.povertyactionlab.org/research-resources/cost-effectiveness>.

americanos utilizando-se a taxa de câmbio padrão para os anos de referência. Como essa análise de custo-efetividade foi realizada dez anos após o ano base da avaliação de impacto, foi considerada a inflação acumulada do período, tendo sido utilizada a taxa de inflação média nos Estados Unidos para o período em questão. O custo total estimado foi de $28.757 (em dólares americanos de 2011).

Para calcular o impacto total do Programa Balsakhi sobre a população considerando o primeiro ano da intervenção, foi feita a multiplicação do impacto médio estimado por criança sobre a nota (0,138 desvio-padrão) pelo número total de crianças em turmas tratadas (6.395 crianças), isto é, crianças em turmas que tiveram acesso à intervenção, mas não necessariamente tendo participado da turma de reforço. Assim, o impacto total estimado foi de 0,138 x 6.395 = 883 desvios-padrão (DP).

A Razão Custo-Efetividade (RCE), que nesse caso mede qual foi o investimento necessário por desvio-padrão adicional de impacto obtido foi calculada da seguinte maneira:

$$RCE = \frac{Custo\ Total}{Impacto} = \frac{\$\ 28.757}{883} = \$32,59$$

A análise de custo-efetividade conduzida pelo J-PAL incluiu ainda uma etapa de análise de sensibilidade, considerando para tal o intervalo de confiança de 90% do impacto estimado. Para o limite superior e para o limite inferior estimados do impacto, foi calculada a RCE de maneira análoga à comentada anteriormente. A tabela a seguir apresenta os resultados, replicando também o resultado associado ao ponto estimado que foi apresentado antes para fins de comparação.

Tabela 3 | Análise de sensibilidade dos resultados da análise de custo-efetividade do Programa Balsakhi

Análise de sensibilidade	Impacto estimado (DP)	Custo por DP adicional
Limite superior	0,215	$20,88
Ponto estimado	0,138	$32,59
Limite inferior	0,061	$74,10

Fonte: adaptado dos resultados apresentados para a análise de custo-efetividade do Programa Balsakhi conduzido pelo Abdul Latif Jameel Poverty Action Lab (J-PAL) com base nos impactos estimados no estudo de Banerjee et al. (2007). Esses resultados e a tabela completa estão disponíveis em: <https://www.povertyactionlab.org/research-resources/cost-effectiveness>.

5.3. Avaliações considerando o custo por unidade de produto ou resultado

No caso de muitas políticas públicas, não estão disponíveis as informações sobre os impactos gerados por elas, podendo isso se dar por inúmeros motivos como, por exemplo, a impossibilidade de mensuração dos impactos (ex: ausência de grupo de controle, falta de recursos para pesquisa de campo) ou intervalo de tempo ainda insuficiente para que se possa observar impactos na prática (ex: efeitos sobre empregabilidade e salários gerados por cursos profissionalizantes não são imediatos).

Nesses casos, pode-se optar por confrontar os custos a produtos ou resultados gerados pela política pública em análise a partir do cálculo de razões análogas à *RCE*, mas nos quais se divide o custo total pelo produto ou resultado gerado (ex: o número de refeições servidas, o número de adultos que concluíram o EJA). Ressalta-se, no entanto, que embora esse tipo de avaliação seja importante para verificar a eficiência dos processos envolvidos na implementação da política, ele não é informativo sobre o cumprimento dos objetivos finalísticos da mesma, uma vez que não trata dos impactos gerados propriamente ditos.

Exemplo 3

Programa Bom Prato

O Programa Bom Prato[55] foi criado no ano de 2000 pelo Governo do Estado de São Paulo e oferece refeições saudáveis e balanceadas a preços populares. O programa tem como objetivo garantir a segurança alimentar de pessoas em situação de vulnerabilidade social, sobretudo na dimensão de renda.

Segundo informações disponíveis na página de descrição do programa[56], em 2018 o programa conta com 53 restaurantes em funcionamento. São servidas duas refeições:

- Café da manhã, ao preço de R$ 0,50 ao usuário. O cardápio é composto por: leite com café, achocola-

[55] Todas as informações apresentadas no presente exemplo foram baseadas na descrição do Programa Bom Prato disponível no site da Secretaria de Estado de Desenvolvimento Social do Governo do Estado de São Paulo (<http://www.desenvolvimentosocial.sp.gov.br/portal.php/bomprato>).

[56] Site: <http://www.desenvolvimentosocial.sp.gov.br/portal.php/bomprato>. Acessado em 22 de novembro de 2018.

tado ou iogurte, pão com margarina, requeijão ou frios e uma fruta da estação;

- Almoço, ao preço de R$ 1,00 ao usuário. O cardápio é composto por: arroz, feijão, salada, legumes, um tipo de carne, farinha de mandioca, pãozinho, suco e sobremesa (geralmente uma fruta da época).

No caso da unidade "Bom Prato 25 de Março", de acordo com o relatório de prestação de contas de 2017 (3 Período, referente a 01/07/2017 a 31/12/2017)[57], considerando o que foi executado de fato, foram servidas 230.220 refeições no total (considerando café da manhã e almoço) e o valor total pago pela Secretaria de Desenvolvimento Social para essa unidade no período foi de R$ 1.155.311,84. Dessa forma, pode-se calcular o custo médio por refeição servida para esse caso específico da seguinte forma:

$$\textit{Custo médio por refeição servida} = \frac{R\$\ 1.155.311,84}{230.220} = R\$\ 5,02$$

Exemplo 4

Prisões gerenciadas por operadores privados

Cabral e Lazzarini (2010) discutem os impactos associados à provisão de serviços de utilidade pública por atores privados, analisando em particular o caso da terceirização de prisões no Paraná no período de 2001 a 2006. Os autores fazem uma análise comparativa entre prisões gerenciadas integralmente pelo Estado e prisões gerenciadas por operadores privados com supervisão pública (terceirizadas), considerando indicadores de custos e de qualidade dos serviços ofertados.

No caso dos indicadores de custos, os autores obtiverem acesso apenas a dados referentes ao ano de 2004. Em uma das análises do artigo, eles calculam o custo médio mensal por detento para cada tipo de administração prisional (geridas pelo Estado ou terceirizadas), sendo que, para isso, dividem o custo mensal total (incluindo remuneração de pessoal, gastos com materiais de consumo, despesas com energia, água e comunicação e despesas com outros serviços) pelo número total de indivíduos custodiados. Os resultados indicam que:

[57] Disponível em: <http://www.desenvolvimentosocial.sp.gov.br/portal.php/bomprato>. Acessado em 22 de novembro de 2018.

- No caso das prisões gerenciadas integralmente pelo Estado, o custo médio mensal estimado foi de R$ 1.387 por interno;
- No caso das prisões gerenciadas por operadores privados com supervisão pública (terceirizadas), o custo médio mensal estimado foi de R$ 1.266 por interno.

Os autores argumentam que, embora não tenha sido possível comparar individualmente cada unidade prisional durante o período analisado no estudo (2001 a 2006), esse exercício provê evidências de que o custo da operação privada é menor, pelo menos considerando esse contexto específico (ano, localidade e itens de custo incluídos na análise).

REFERÊNCIAS

ANGRIST, J.D.; PISCHKE, J.S. Mostly harmless econometrics: An empiricist's companion. Princeton University Press, 2008.

AOS, S., LIEB, R., MAYFIELD, J., MILLER, M., PENNUCCI, A. Benefits and costs of prevention and early intervention programs for youth. Washington State Institute for Public Policy, 2004.

ARVATE, P., FALSETE, F.O., RIBEIRO, F.G.; SOUZA, A.P. Lighting and Homicides: Evaluating the Effect of an Electrification Policy in Rural Brazil on Violent Crime Reduction. Journal of Quantitative Criminology, p. 1-32, 2017.

BANERJEE, A. et al. Can e-governance reduce capture of public programs. Experimental evidence from a financial reform of India's employment guarantee, 2014. Disponível em: <http://www.3ieimpact.org/media/filer_public/2015/07/28/ie_31-_can_e-governance_reduce_capture_of_public_works_programme.pdf>.

BANERJEE, A.V., COLE, S., DUFLO, E.; LINDEN, L. Remedying education: Evidence from two randomized experiments in India. The Quarterly Journal of Economics, 122(3), pp.1235-1264, 2007.

BASCH, C.E. et al. Avoiding type III errors in health education program evaluations: a case study. Health Education Quarterly, v. 12, n. 3, p. 315-331, 1985.

BATTISTIN, E.; RETTORE, E. Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. Journal of Econometrics, 142(2), pp.715-730, 2008.

MORETTIN, P.A.; BUSSAB, W.O. Estatística básica. Editora Saraiva, 9ª Edição, 2017.

CABRAL, S.; LAZZARINI, S.G. Impactos da participação privada no sistema prisional: evidências a partir da terceirização de prisões no Paraná. Revista de Administração Contemporânea — RAC, v. 14, n. 3, art. 1, p. 395-413, 2010.

CALIENDO, M.; KOPEINIG, S. Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22(1), pp.31-72, 2008.

CAMELO, R.S., TAVARES, P.A.; SAIANI, C.C.S. Alimentação, nutrição e saúde em programas de transferência de renda: evidências para o Programa Bolsa Família. Revista EconomiA, 2009.

CHILE. Dirección de Presupuestos do Ministerio de Hacienda. Evaluación Ex-Post: Conceptos y Metodologías. Santiago, 2015. Disponível em: <http://www.dipres.gob.cl/598/articles-135135_doc_pdf.pdf>.

CHILE. Dirección de Presupuestos do Ministerio de Hacienda. Metodología Para La Elaboración De Matriz De Marco Lógico. Santiago, 2014. Disponível em: <http://siare.clad.org/siare/innotend/evaluacion/chile2/m2004.pdf>.

DHALIWAL, I., DUFLO, E., GLENNERSTER, R.; TULLOCH, C. Comparative cost-effectiveness analysis to inform policy in developing countries: a general framework with applications for education. Education Policy in Developing Countries, pp.285-338, 2013.

DJIMEU, E.W.; HOUNDOLO, D.G. Power calculation for causal inference in social science: sample size and minimum detectable effect determination. Journal of Development Effectiveness, 8(4), pp.508-527, 2016.

DUFLO, E. et al. A wide angle view of learning: Evaluation of the CCE and LEP programmes in Haryana, India. 3ie Impact Evaluation Report, v. 22, 2015. Disponível em: <http://www.3ieimpact.org/media/filer_public/2015/02/24/ie_22_evaluation_of_cce_and_lep_in_haryana.pdf>.

DUFLO, E., GLENNERSTER, R.; KREMER, M. Using randomization in development economics research: A toolkit. Handbook of Development Economics, 4, pp.3895-3962, 2007.

ESPÍRITO SANTO. Lei n. 10.744, de 05 de outubro de 2017. Institui o Sistema de Monitoramento e de Avaliação de Políticas Públicas do Espírito Santo. Vitória, 2017. Disponível em: <http://www.al.es.gov.br/antigo_portal_ales/images/leis/html/LEI107442017.html>. Acesso em: 11 jun. 2018.

GERTLER, P. Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. American Economic Review, 94(2), pp.336-341, 2004.

GERTLER, P.J., MARTINEZ, S., PREMAND, P.; RAWLINGS, L.B. Avaliação de Impacto na Prática, 2ª Edição. World Bank Publications, 2018.

H M TREASURY. UK Government. The Magenta Book: Guidance for evaluation. London, 2011.

IJSN – INSTITUTO JONES DOS SANTOS NEVES. Sistema de Monitoramento e Avaliação de Políticas Públicas do Estado do Espírito Santo (SiMAPP). Vitória, 2018. (Nota Técnica, n. 56). Disponível em: <http://www.ijsn.es.gov.br/component/attachments/download/6376>. Acesso em: 31 out. 2018.

J-PAL - ABDUL LATIF JAMEEL POVERTY ACTION LAB. J-PAL Costing Guidelines. 2016. Disponível em: <https://www.povertyactionlab.org/research-resources/cost-effectiveness>.

J-PAL - ABDUL LATIF JAMEEL POVERTY ACTION LAB. Introduction to Evaluations. 2016. Disponível em: <https://www.povertyactionlab.org/sites/default/files/resources/Introduction-to-Evaluations.pdf>.

LEME, M.C., LOUZANO, P., PONCZEK, V.; SOUZA, A.P. The impact of structured teaching methods on the quality of education in Brazil. Economics of Education Review, 31(5), pp.850-860, 2012.

MENEZES FILHO, N.; PINTO, C.C.X. (organizadores). Avaliação Econômica de Projetos Sociais, 3ª Edição. Fundação Itaú Social, 2017.

MÉXICO. CONSEJO NACIONAL DE EVALUACIÓN DE LA POLÍTICA DE DESARROLLO SOCIAL (CONEVAL). Modelos de términos de referência para la Evaluación de Procesos de Programas de Desarollo Social. 2017. Disponível em: <https://www.coneval.org.mx/Evaluacion/MDE/Documents/TDR_Procesos.pdf>.

NEWCOMER, K.E., HATRY, H.P.; WHOLEY, J.S. Handbook of practical program evaluation. John Wiley & Sons, 2015.

ROSENBAUM, P.R.; RUBIN, D.B. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), pp.41-55, 1983.

RUBIN, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5), 688, 1974.

ROSSI, P.H.; LIPSEY, M.W.; FREEMAN, H.E. Evaluation: A systematic approach. Sage publications, 2003.

SAUNDERS, R.P.; EVANS, M.H.; JOSHI, P. Developing a process-evaluation plan for assessing health promotion program implementation: a how-to guide. Health promotion practice, v. 6, n. 2, p. 134-147, 2005.

SKOUFIAS, E., DAVIS, B.; BEHRMAN, J.R. An evaluation of the selection of beneficiary households in the education, health, and nutrition program (PROGRESA) of Mexico. International Food Policy Research Institute, Washington, DC, 1999.

TAVARES, P.A. The impact of school management practices on educational performance: Evidence from public schools in São Paulo. Economics of Education Review, 48, pp.1-15, 2015.

VERMEERSCH, C., ROTHENBUHLER, E.; STURDY, J. Impact evaluation toolkit: measuring the impact of results-based financing on maternal and child health. World Bank, Washington, DC, 2012.