TESTAR, APRENDER, ADAPTAR: DESENVOLVER AS POLÍTICAS PÚBLICAS MEDIANTE EXPERIMENTOS ALEATÓRIOS CONTROLADOS*

Laura Haynes**
Owain Service***
Ben Goldacre****
David Torgerson*****

1 INTRODUÇÃO

Os experimentos aleatórios controlados (EACs) são a melhor forma de determinar se uma política funciona. Têm sido usados por mais de sessenta anos para comparar a eficiência de novos medicamentos.[1] Os EACs são cada vez mais usados nos programas de desenvolvimento internacional, para comparar a eficiência de custo de diferentes intervenções para reduzir a pobreza (Banerjee e Duflo, 2011; Karlan e Appel, 2011). São também empregados amplamente por companhias que desejam saber qual *layout* de *website* gera mais vendas. Entretanto, ainda não se tornaram prática comum na maioria das áreas de políticas públicas (gráfico 1).

Este trabalho argumenta que se deve e pode usar EACs muito mais extensamente em política pública interna para testar a eficiência das intervenções novas e existentes e suas variações; aprender o que funciona e o que não; e adaptar políticas de tal forma que melhorem de forma constante e evoluam tanto em termos de qualidade como de eficiência.

* Os autores agradecem aos departamentos do governo por compartilharem sua recente pesquisa envolvendo experiências de campo. Agradecem também aos professores Peter John, Rachel Glennester, Don Green, David Halpern e outros membros do Behavioural Insights Team por seus comentários a este trabalho, e também a Michael Sanders, por editar este artigo.

Nota dos editores: este texto foi publicado originalmente em inglês com o título *Test, learn, adapt: developing public policy with randomised controlled trials*, em 2012, e está disponível em: <http://goo.gl/y3Gy8E>. Os autores gentilmente autorizaram a publicação em língua portuguesa.

** Chefe de Pesquisa de Políticas no Behavioural Insights Team e pesquisadora visitante no King's College, em Londres.

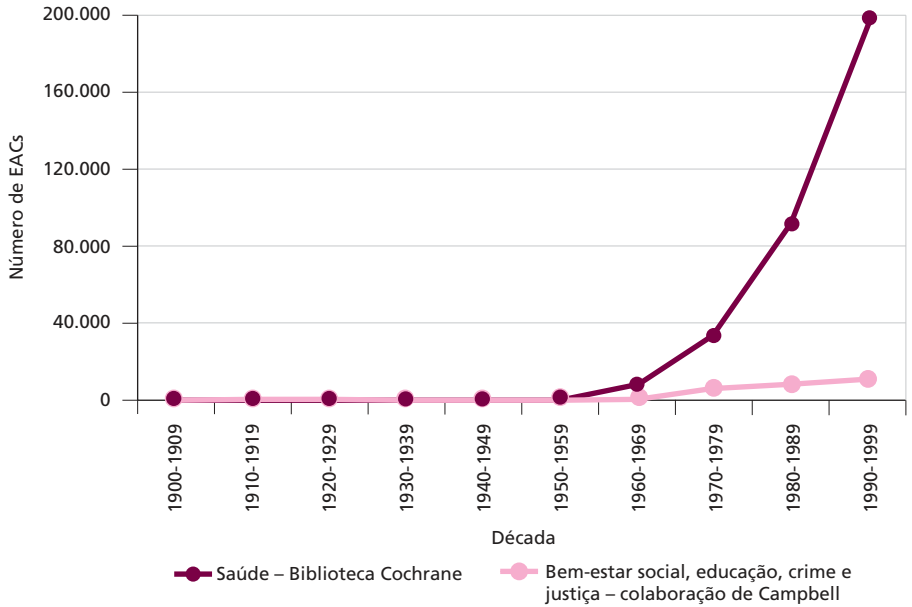
*** Vice-diretor do Behavioural Insights Team.

**** Pesquisador sênior na Escola de Higiene e Medicina Tropical de Londres.

***** Diretor da Unidade de Experimentos de York.

1. O primeiro EAC publicado na medicina é atribuído a Sir Brandford Hill, um epidemiologista do Medical Research Council da Inglaterra. O experimento, publicado no *British medical journal* em 1948, testou se a estreptomicina era efetiva no tratamento da tuberculose.

GRÁFICO 1
EACs no século XX em saúde e em bem-estar social, educação, crime e justiça



Fonte: Shepherd (2007).

A seção 2 deste trabalho estabelece o que é um EAC e por que este é importante. O texto aborda muitos dos argumentos contrários ao uso de EACs em política pública e defende que os experimentos não são tão difíceis de serem feitos como frequentemente se presume, podendo ter ótimas relações custo-benefício para avaliação de resultados de políticas e para estimar o retorno dos dispêndios.

A seção 3 do trabalho destaca nove passos-chave que qualquer EAC precisa ter. Muitos destes passos são fundamentais para qualquer política, outros necessitarão de apoio de acadêmicos ou centros especializados no governo.

A filosofia de "testar, aprender, adaptar" definida neste trabalho está no âmago da forma na qual o Behavioural Insights Team trabalha.[2] A abordagem "testar, aprender, adaptar" tem potencial para ser usada em quase todos os aspectos da política pública.

1) *Testar* uma intervenção significa assegurar-se de que foram implantadas medidas robustas que possibilitam avaliar sua eficiência.

2. Nota dos revisores técnicos: o Behavioural Insights Team é a equipe do Gabinete do Governo Britânico dedicada à aplicação dos instrumentos da economia comportamental e da psicologia às políticas e serviços públicos.

2) *Aprender* consiste em analisar o resultado da intervenção, de forma a que se possa identificar o que funciona e se a dimensão do efeito é ou não suficientemente grande para estabelecer uma boa relação custo-benefício.

3) *Adaptar* significa usar essa aprendizagem para modificar a intervenção (se necessário) de tal maneira que se refine continuamente a forma na qual a política é projetada e implementada.

2 O QUE SÃO EACs E POR QUE SÃO IMPORTANTES?

2.1 O que é um experimento aleatório controlado?

Muitas vezes, é preciso saber qual de duas ou mais intervenções é a mais eficaz para atingir um resultado específico e mensurável. Por exemplo, quando se quer comparar uma nova intervenção com a prática corrente, ou comparar níveis diferentes de "dosagem" entre si (como consultas domiciliares a uma adolescente grávida uma vez por semana, ou duas vezes por semana).

Convencionalmente, para avaliar se uma intervenção tem um benefício, basta implementá-la e observar os resultados. Por exemplo, é possível estabelecer um programa intensivo de assistência "de volta ao trabalho" e monitorar se os participantes saíram do seguro-desemprego com mais rapidez que antes do programa ser introduzido.

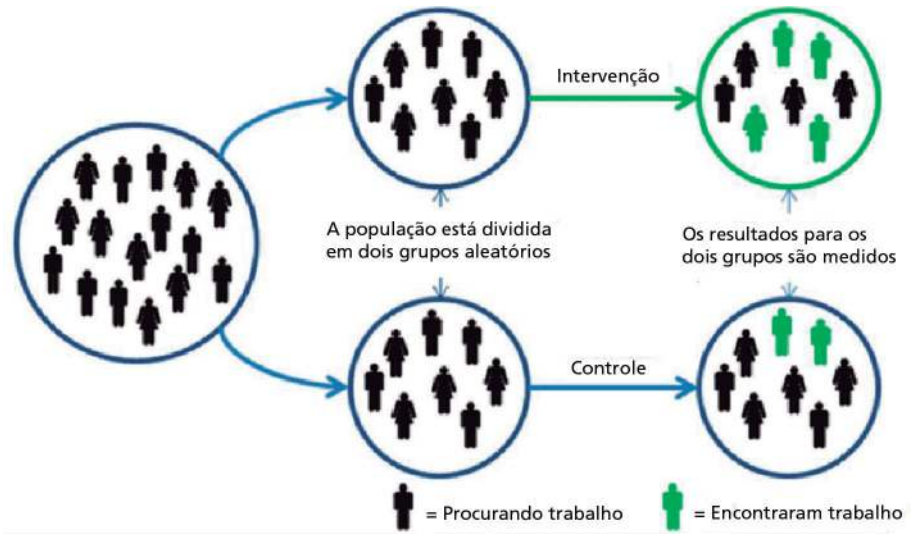
Entretanto, essa abordagem tem diversas desvantagens que tornam difícil identificar se foi a intervenção que teve o efeito ou algum outro fator, principalmente os fatores não controlados, externos. Se houver crescimento econômico, por exemplo, podemos esperar que mais pessoas encontrem emprego independentemente de uma nova política.

Outro desafio analítico complicado trata do chamado "viés de seleção"; ou seja, as pessoas que querem participar do programa de volta ao trabalho são sistematicamente diferentes daquelas que não querem. Elas talvez sejam mais motivadas a encontrar trabalho, significando que os benefícios da nova intervenção serão superestimados. Existem técnicas estatísticas para tentar controlar diferenças preexistentes entre os grupos que recebem intervenções diferentes, mas estas são sempre imperfeitas e podem introduzir mais vieses.

Os experimentos aleatórios controlados contornam esse problema, ao assegurar que os indivíduos, ou os grupos de pessoas, que recebem ambas as intervenções sejam o máximo possível parecidos. No exemplo aqui utilizado, de programa "de volta ao trabalho", isto pode envolver a identificação de 2 mil pessoas que seriam elegíveis para o novo programa e a divisão aleatória delas em dois grupos de 1 mil, dos quais um receberia a intervenção corrente e o outro receberia a nova intervenção.

Ao designar aleatoriamente as pessoas aos grupos, pode-se eliminar a possibilidade de fatores externos afetarem os resultados e demonstrar que quaisquer diferenças entre os dois grupos são exclusivamente resultado da diferença no tratamento que receberem.

FIGURA 1
Ilustração de um EAC para testar um novo programa "de volta ao trabalho" (resultado positivo)

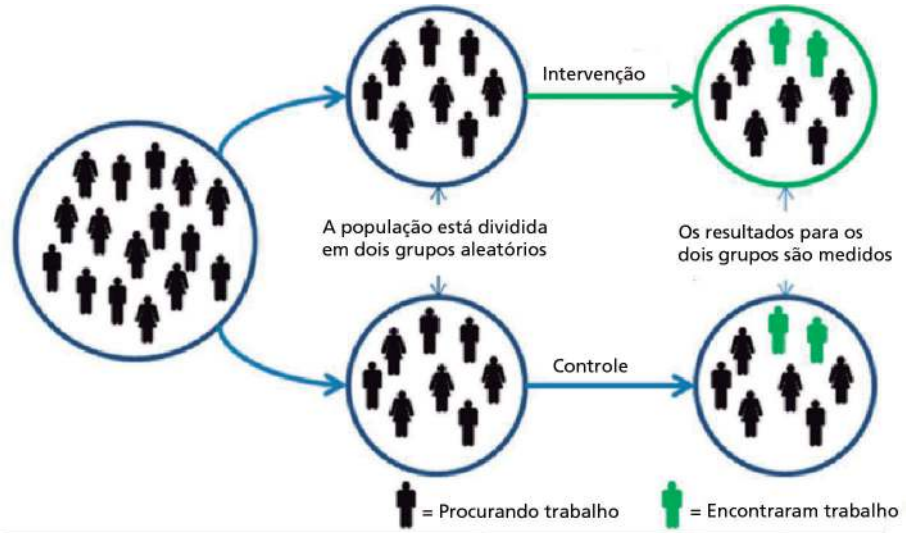


Elaboração dos autores.

A seção 3 deste artigo descreve mais detalhadamente como realizar um experimento aleatório controlado, mas no âmago de qualquer EAC existem diversos elementos-chave. Os EACs operam ao dividir a população em dois ou mais grupos aleatoriamente, dando uma intervenção a um grupo, outra a um outro, e mensurando o resultado predeterminado para cada grupo. Este processo está resumido na figura 1.

Suponha-se que se tenha testado um novo programa "de volta ao trabalho" que visa ajudar os que buscam emprego. A população avaliada é dividida aleatoriamente em dois grupos, e apenas um destes grupos recebe a nova intervenção ("o grupo de tratamento") – neste caso, o programa "de volta ao trabalho". O outro grupo ("o grupo de controle") recebe o apoio usual que uma pessoa que busca por emprego teria. Neste caso, o grupo de controle é equivalente ao grupo que recebe o placebo em um experimento de medicamento.

FIGURA 2
Ilustração de um EAC para testar um novo programa "de volta ao trabalho" (resultado neutro)



Elaboração dos autores.

No exemplo da figura 1, os candidatos que encontraram trabalho em tempo integral seis meses antes do início do experimento estão na cor verde. O experimento mostra que o número de indivíduos no novo programa "de volta ao trabalho" que estão agora trabalhando é muito maior que os do grupo de controle.

É importante observar que duas figuras no grupo de controle também encontraram trabalho, talvez tendo se beneficiado dos serviços usuais de apoio a quem está no seguro-desemprego e procura emprego.

Se o novo programa "de volta ao trabalho" não fosse melhor que o serviço corrente prestado aos que procuram emprego, seria possível observar um padrão similar tanto no grupo de tratamento como no grupo de controle. Isto está ilustrado na figura 2, que mostra um conjunto diferente de resultados para o programa.

No caso, os resultados do experimento demonstram que o novo e caro programa "de volta ao trabalho" não é melhor que a prática corrente. Se não houvesse um grupo de controle, o resultado mostraria pessoas conseguindo empregos após participarem do novo programa "de volta ao trabalho" e concluído erroneamente que isto se devia ao programa. Fato que poderia ter levado a lançar o novo e dispendioso (e ineficaz) tratamento. Um erro como este foi evitado pelo Departamento de Trabalho e Pensões (DWP – em inglês, *for Work and Pensions*) em um EAC real sobre o custo-benefício de diferentes tipos de intervenções (box 1)

BOX 1
Usando EACs para saber o que realmente funciona para levar as pessoas de volta ao trabalho

Em 2003, o DWP[1] realizou um EAC para examinar o impacto de três novos programas sobre pedido de benefícios de invalidez: apoio no trabalho, apoio focado nas necessidades individuais de saúde, ou ambos (DWP, 2006a).[2] O apoio extra custou £1.400,00 em média, mas a experiência não encontrou nenhum benefício além do apoio padrão que já estava disponível. O EAC economizou para o contribuinte muitos milhões de libras, pois forneceu evidência inequívoca de que o dispendioso apoio adicional não teve o efeito pretendido.

Mais recentemente, o DWP desejava investigar se a frequência da entrevista de revisão exigida para os desempregados que recebem seguro-desemprego poderia ser reduzida sem piorar os resultados.

Em um experimento envolvendo mais de 60 mil pessoas, o processo de entrevista de revisão do benefício quinzenal tradicional foi comparado a diversos outros menos exigentes (por exemplo, entrevista pelo telefone, com menor frequência). Todas as alternativas ao *status quo* testadas em experimentos suficientemente grandes para mostrar efeitos confiáveis aumentaram o tempo que as pessoas gastam para encontrar emprego (DWP, 2006b). Como resultado, apesar de outras mudanças no sistema de benefícios, a política do DWP continua exigindo a realização de entrevistas frequentes.

Elaboração dos autores.
Notas: [1] Nota dos revisores técnicos: O DWP britânico equivale aos ministérios do Trabalho e da Previdência brasileiros.
[2] Esse é um projeto de interação que permite a determinação dos efeitos separados e combinados de duas intervenções. Estes projetos são especialmente úteis nas situações em que existem questões sobre o efeito adicional de uma/ou mais características de um programa complexo.

Sempre que houver potencial para fatores externos afetarem os resultados de uma política, vale a pena considerar usar EACs para testar a eficiência do tratamento antes de implementá-la em toda a população. Quando não se procede desta maneira, é fácil confundir mudanças que poderiam ter ocorrido de qualquer forma com o impacto de um determinado tratamento.

No exemplo fictício "de volta ao trabalho" aqui sugerido, assume-se o interesse principal de entender qual das duas intervenções de larga escala funciona com mais eficiência. Em muitos casos, um EAC pode se voltar a temas que vão além da política principal, podendo ser usado para comparar diversas formas de implementar aspectos menores da mesma política.

Como muitos outros exemplos que serão dados ao longo deste texto mostram (box 2), uma das vantagens dos experimentos aleatórios controlados é que eles também permitem que se teste a eficácia de determinados aspectos de um programa mais amplo. Testar pequenas partes de um programa possibilita que os formuladores de políticas refinem continuamente a política, aprimorando o aspecto particular do tratamento que tenha o maior impacto.

BOX 2

Demonstração do impacto de mensagens de texto nos reembolsos de multas

O Serviço de Tribunais e o Behavioural Insights Team queriam testar se enviar mensagens de texto para pessoas que deixaram de pagar suas multas de tribunal poderia incentivá-las ou não a pagar antes de enviar um oficial de justiça às suas residências. A forma como esta questão foi respondida é um exemplo claro da abordagem "testar, aprender, adaptar" e mostra que o teste simultâneo de diversas variações pode ajudar a descobrir o que funciona melhor.

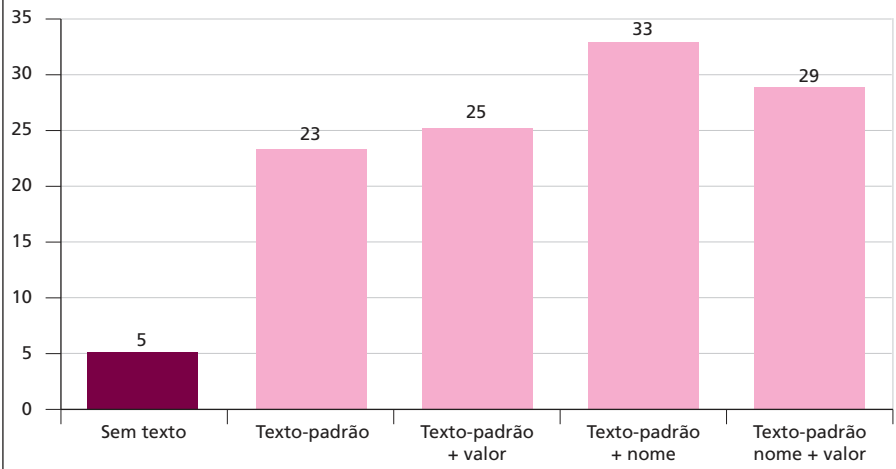
Na experiência inicial, os indivíduos foram alocados aleatoriamente em cinco grupos diferentes. Alguns não receberam nenhuma mensagem (grupo de controle), enquanto outros (grupos de tratamento) receberam um texto-padrão de lembrança ou uma mensagem mais personalizada (incluindo o nome do recipiente, o valor devido, ou ambos).

A experiência mostrou que os lembretes por mensagem podem ser altamente eficazes (gráfico 2).

GRÁFICO 2

Experiência inicial: índices de reembolso por indivíduos (N=1.054)[1]

(Em %)



Fonte: Behavioural Insights Team.

Nota: [1] Os números refletem os índices de resposta às mensagens de texto que foram enviadas (o Serviço de Tribunais possuía números corretos dos celulares).

Um segundo experimento foi realizado usando uma amostra maior (N=3.633) para determinar que aspectos das mensagens personalizadas poderiam contribuir para aumentar os índices de pagamento. O padrão de resultados foi muito similar ao primeiro experimento. Entretanto, o segundo permitiu assegurar não apenas que as pessoas tinham mais probabilidade de pagar sua multa vencida se recebessem uma mensagem de texto contendo seu nome, mas também que o valor médio dos reembolsos de multa subiu mais de 30%.

(Continua)

(Continuação)

Os dois experimentos foram realizados com custo muito baixo: como os dados do resultado já estavam sendo coletados pelo Serviço de Tribunais, o único custo foi o tempo para membros da equipe montarem o experimento. Se tivesse sido lançada em nível nacional, os lembretes por mensagem de texto melhorariam a cobrança de multas não pagas – estima-se que simplesmente enviar um texto personalizado em vez de um padronizado produza mais de £ 3 milhões anualmente. A economia obtida com os textos personalizados é muitas vezes maior que não enviar nenhum lembrete por mensagem de texto. Além desta economia financeira, o Serviço de Tribunais estima que enviar lembretes em texto personalizado pode reduzir a necessidade de até 150 mil intervenções de oficiais de justiça anualmente.

Elaboração dos autores.

Independentemente de se comparar duas intervenções de grande escala ou detalhes de uma única política, os mesmos princípios básicos de um EAC são válidos; ao compararem-se dois grupos idênticos escolhidos aleatoriamente, pode-se controlar uma grande variedade de fatores que permitem entender o que funciona e o que não funciona.

BOX 3
O elo entre teorias de crescimento, inovação e EACs

Os dois experimentos foram realizados com custo muito baixo: como os dados do resultado já estavam sendo coletados pelo Serviço de Tribunais, o único custo foi o tempo para membros da equipe montarem o experimento. Se tivesse sido lançada em nível nacional, os lembretes por mensagem de texto melhorariam a cobrança de multas não pagas – estima-se que simplesmente enviar um texto personalizado em vez de padronizado produza mais de £ 3 milhões anualmente. A economia obtida com os textos personalizados é muitas vezes maior que não enviar nenhum lembrete por mensagem de texto. Além desta economia financeira, o Serviço de Tribunais estima que enviar lembretes em texto personalizado pode reduzir a necessidade de até 150 mil intervenções de oficiais de justiça anualmente.

O crescente interesse no uso de EACs como uma ferramenta importante de elaboração e implementação de políticas ressoa com correntes mais amplas de pensamento. Quando o dinheiro é escasso, é essencial se certificar de que ele está sendo gasto em abordagens que funcionam; e mesmo melhorias marginais em eficácia de custo são preciosas. Os EACs são uma ferramenta extremamente poderosa para identificar boas relações de custo-eficácia e evitar o gasto de baixo retorno.

(Continua)

(Continuação)

Esses métodos também ressoam fortemente com pontos de vista emergentes sobre progresso social e econômico. Muitos pensadores de destaque concluíram que, em sistemas complexos, desde ecossistemas biológicos até economias modernas, boa parte do progresso – se não a maioria – ocorre através de um processo de tentativa e erro (Harford, 2011). Economias e ecossistemas que se tornam demasiadamente dominados por um estreito espectro de práticas, espécies ou empresas são mais vulneráveis que sistemas mais diversos (Taleb, 2007; Christensen, 2003). Similarmente, estes pensadores tendem a ser céticos sobre a capacidade de mesmo os especialistas e líderes mais sábios oferecerem estratégia abrangente ou plano que detalhe a "melhor" prática ou resposta real (certamente com abrangência universal). Em vez disso, encorajam o cultivo deliberado da diversidade em sistemas que eliminem as variações menos eficazes e compensem e reproduzam as variedades que tenham melhor desempenho.

A expressão prática desse pensamento inclui o impulso para maior descentralização de formulação de políticas e o uso de mercados para fornecimento de mercadorias e serviços. O incentivo à variação precisa ser combinado a mecanismos que identifiquem e favoreçam inovações bem-sucedidas. Isto inclui aumentar a transparência e a retroalimentação nos mercados de bens e nos serviços públicos, observando que estes levam à expansão seletiva dos melhores ofertantes e, muitas vezes, ao crescimento de ofertantes menores e independentes (Luca, 2011). Nos serviços públicos, e quando os mercados e o pagamento por resultados podem ser inapropriados, os EACs e os experimentos com muitos grupos podem desempenhar um papel poderoso, especialmente quando estes resultados são amplamente informados e aplicados.

Elaboração dos autores.

2.2 Uma defesa dos EACs: desmascarando alguns mitos

Existem muitas áreas nas quais os experimentos aleatórios são prática comum, e onde deixar de fazê-los seria considerado bizarro ou mesmo negligente. Os EACs são o meio universal de avaliar qual dos dois tratamentos médicos funciona melhor – se é um novo medicamento comparado ao melhor tratamento atual, duas formas diferentes de cirurgia de câncer, ou até mesmo duas meias de compressão diferentes. Nem sempre foi assim: quando os experimentos foram introduzidos na medicina, sofreram forte resistência de alguns médicos, muitos dos quais acreditavam que seu julgamento pessoal e especializado era suficiente para decidir se um determinado tratamento era eficiente.

Os EACs são cada vez mais usados para investigar a eficácia e o retorno de diversos programas de desenvolvimento (box 4).

BOX 4
O uso de EACs para melhorar resultados eduacionais na Índia

Uma das áreas de rápido crescimento no uso de EACs em anos recentes deu-se na área de projetos de desenvolvimento internacional. Numerosos experimentos foram conduzidos para determinar como reduzir a pobreza no mundo em desenvolvimento, como incentivar variedades agrícolas de baixo rendimento, encorajar o uso de mosquiteiros, garantir que os professores compareçam às aulas, promover o empreendedorismo e aumentar as taxas de vacinação.

Por exemplo, o esforço, nas décadas recentes, de universalizar a educação nos países em desenvolvimento levou ao aumento das matrículas e do comparecimento escolar. Contudo, a qualidade da educação das crianças de lares pobres ainda continua problemática: em 2005, uma pesquisa na Índia indicou que mais de 40% das crianças menores de 12 anos não conseguia ler um parágrafo simples e 50% não conseguia fazer uma subtração simples.

Em parceria com uma ONG de educação, os pesquisadores conduziram um RCT para determinar se um programa de apoio educacional de baixo custo melhoraria os resultados educacionais na Índia. Mais de duzentas escolas foram escolhidas aleatoriamente para receber um tutor para a terceira ou quarta série. O impacto do programa foi estimado pela comparação das notas da terceira série destas escolas com as demais.

Os tutores eram mulheres da comunidade local, que recebiam uma fração do salário do professor e trabalhavam por meio turno, com grupos de alunos com desempenho inferior em relação a seus colegas. Os resultados indicaram que o programa de apoio melhorou significativamente as notas dos alunos, especialmente em matemática. O programa foi considerado tão bem-sucedido (e eficiente em termos de custo quando comparado a outros programas para melhorar o desempenho escolar) que foi aplicado ao restante da Índia (Banerjee *et al*., 2007).

Elaboração dos autores.

Nos negócios, quando as companhias querem descobrir qual é o melhor *design* para uma página na internet, é comum mostrar as diversas opções para os visitantes e, então, rastrear seus cliques e seu comportamento de compra (box 5).

BOX 5
O uso de EACs para melhorar o desempenho empresarial

Muitas companhias usam EACs para testar as reações do consumidor a diferentes apresentações de seus produtos *on-line*. Pouca informação é pública, mas é conhecido que companhias como Amazon e eBay usam o tráfego em seus *sites* para testar o que funciona melhor para impulsionar as compras. Por exemplo, alguns clientes veem uma determinada configuração de uma página, enquanto outros veem uma diferente. Ao rastrear os cliques e o comportamento de compra de clientes que veem as diferentes versões do *site*, as companhias podem ajustar o desenho da página para maximizar os lucros. Alguns exemplos são fornecidos a seguir.

(Continua)

(Continuação)

Durante a iniciativa recente de coletar fundos para a Wikipedia, uma fotografia do fundador, Jimmy Wales apareceu nas propagandas das doações na parte superior da página: isto foi o resultado de uma série de experimentos comparando formas diferentes de publicidade, transmitindo-as aleatoriamente para os visitantes do *site* e monitorando se estes doaram ou não.

Netflix é uma companhia que oferece filmes *on-line* e que executa diversos experimentos com o usuário. Quando experimentavam a "Sala de Projeção Netflix", uma nova forma de ver antecipadamente filmes, produziram quatro versões diferentes do serviço. Estes foram mostrados para quatro grupos de 20 mil assinantes, e um grupo de controle recebeu o serviço normal do Netflix. Os usuários foram então monitorados para verificar se assistiram a mais filmes (Davenport e Harris, 2007).

A Delta Airlines também usou o experimento para melhorar seu *website*. Em 2006, apesar de números crescentes de pessoas estarem reservando suas passagens *on-line*, o tráfego para o *website* da Delta Airlines não gerava o número esperado de reservas. Quase 50% dos visitantes do *site* saíam antes de concluir o processo de reserva: após selecionarem seu voo, os prováveis clientes geralmente abandonavam a reserva quando chegavam à página que pedia a inserção de suas informações pessoais (nome, endereço e detalhes do cartão de crédito).

Em vez de mudar todo o *site*, a Delta se concentrou em fazer mudanças nas páginas específicas que deixavam de converter possíveis clientes em vendas. Diversas variações foram testadas *on-line*, direcionando aleatoriamente clientes para versões diferentes das páginas. A Delta descobriu que, ao remover as instruções detalhadas na parte superior da página que pede que o cliente insira suas informações pessoais, estes tiveram propensão maior a completar a compra. Como resultado da implementação desta e de outras mudanças sutis no *site*, identificadas durante o processo de testes, as taxas de conversão para vendas de passagens melhoraram em 5% (Delta Airlines, 2007), uma mudança pequena, mas altamente valiosa.

Elaboração dos autores.

Mas, apesar de haver alguns bons exemplos de formuladores de políticas usando EACs no Reino Unido, eles ainda não são amplamente usados. Isto pode ser parcialmente devido a uma falta de conscientização, mas existem também muitos mal-entendidos sobre EACs, o que os leva a serem inapropriadamente rejeitados.

A seguir, serão abordados cada um desses mitos, tratando das suposições incorretas de que os EACs são sempre difíceis, dispendiosos, antiéticos ou desnecessários. Argumenta-se o perigo da confiança demasiada ao assumir que as intervenções são eficazes, e que os EACs desempenham um papel vital na demonstração não apenas da eficiência de um tratamento, como também do seu retorno.

2.2.1 Não sabemos necessariamente "o que funciona"

Os formuladores de políticas e profissionais geralmente acham que têm um bom entendimento de quais tratamentos podem funcionar e usam estas crenças para

planejar a política. Mesmo se houver bons motivos para acreditar que uma política será eficaz, vale a pena executar um EAC para quantificar o benefício da maneira mais precisa possível. Um experimento pode, ainda, ajudar a demonstrar quais aspectos de um programa estão tendo maior efeito, e como pode este ser melhorado. Por exemplo, ao implementar um novo programa para o financiamento de empreendedores, seria útil saber se, dobrando o valor do dos recursos, haveria um efeito significativo sobre o sucesso, ou se não faz diferença.

Devemos também reconhecer que as previsões confiantes sobre política feitas por especialistas frequentemente se mostram incorretas. Os EACs têm demonstrado que intervenções que foram projetadas para serem eficazes na realidade não o foram (box 1). Também mostraram que intervenções sobre as quais houve ceticismo inicial foram proveitosas. Por exemplo, quando o Behavioural Insights Team e o Serviço de Tribunais avaliaram se as mensagens de texto poderiam incentivar as pessoas a pagarem suas multas de tribunal, poucos previram que um texto personalizado aumentaria tão significativamente os índices e os valores de reembolso (box 2).

Mas existem ainda incontáveis exemplos de EACs que subverteram as suposições tradicionais sobre o que funciona e mostraram que as intervenções tidas como sendo eficazes eram, na realidade, prejudiciais. O caso da aplicação de esteroides (box 6) é um exemplo poderoso de como suposições aparentemente sólidas não são verdadeiras quando testadas. Da mesma forma, o programa Scared Straight, que expõe jovens às realidades de uma vida de crime, é um bom exemplo de uma intervenção política bem intencionada, com uma base de evidência aparentemente sólida, mas que os EACs mostraram ter efeitos adversos (box 7). Os EACs são o melhor método para evitar estes erros, ao oferecer aos formuladores de políticas e profissionais uma evidência robusta da eficácia de uma política, e assegurar que se sabe o que teria acontecido se não houvesse intervenção.

BOX 6
Esteroides para lesões na cabeça: salvando vidas ou matando pessoas?

Durante diversas décadas, os adultos com lesões graves na cabeça eram tratados usando-se injeções de esteroide. Isto fazia sentido em princípio: esteroides reduzem o inchaço, e acreditava-se que o inchaço, ao pressionar o cérebro, dentro do crânio, matava pessoas com ferimentos na cabeça. Entretanto, estas hipóteses não tinham sido submetidas a testes apropriados.

Então, há uma década, essa suposição foi testada em um experimento aleatório (Edwards *et al*., 2005). O estudo foi controverso, e muitos se opuseram a ele, porque achavam que já sabiam que os esteroides eram eficazes. Na realidade, quando os resultados foram publicados, em 2005, revelaram que as pessoas que recebiam injeções de esteroide tinham mais probabilidade de morrer: este tratamento de rotina tinha matado pessoas, e em grandes quantidades, porque os ferimentos na cabeça são frequentes. Os resultados foram tão extremos que o experimento teve de ser suspenso, para evitar dano adicional.

(Continua)

(Continuação)

Esse é um exemplo particularmente dramático de por que os testes sérios de intervenções novas são importantes: sem eles, podemos causar danos não intencionais, sem sequer conhecê-los; e quando novas intervenções se tornam uma prática comum sem boa evidência, então, pode haver resistência a testá-las no futuro.

Elaboração dos autores.

BOX 7
O Programa Scared Straight: dissuadindo delinquentes juvenis ou incentivando-os?

Scared Straight é um programa desenvolvido nos Estados Unidos para deter os delinquentes juvenis e crianças em risco de um comportamento criminoso. O programa expôs crianças à realidade de uma vida de crime, fazendo-as interagir com criminosos encarcerados.

A teoria era de que essas crianças teriam menos probabilidade de se envolver em comportamentos criminosos se tivessem consciência das consequências. Diversos estudos anteriores, que analisaram comportamentos criminosos de participantes antes e após o programa, pareciam apoiar estas hipóteses (Finckenauer, 1982). Os índices de sucesso foram informados como sendo de 94%, e o programa foi adotado em diversos países, inclusive no Reino Unido.

Nenhuma dessas avaliações teve um grupo de controle mostrando o que teria acontecido a esses participantes se não tivessem participado do programa. Diversos EACs foram feitos para corrigir este problema. Uma meta-análise de sete experimentos nos Estados Unidos, que escolheram aleatoriamente metade da amostragem de crianças em risco para o programa, descobriram que o Scared Straight, na realidade, levou a índices mais elevados de comportamento delinquente: "não fazer nada teria sido melhor que expor os jovens ao programa" (Petrosino, Turpin-Petrosino e Buehler, 2003). Análise recente sugere que os custos associados ao programa (amplamente relacionados ao aumento nos índices de reincidência) foram trinta vezes mais elevados que os benefícios, significando que o programa Scared Straight custa para o contribuinte um montante significativo de dinheiro e gera mais crimes (The Social Research Unit, 2012).

Elaboração dos autores.

2.2.2 EACs não precisam ser caros

Os custos de um EAC dependem de como ele é projetado: com planejamento, podem ser mais baratos que outras formas de avaliação. Isto é especialmente verdadeiro quando um serviço já é provido, e quando os dados do resultado são coletados pelos sistemas rotineiros de monitoramento, como em muitas partes do setor público. Em contraste com experimentos médicos, um experimento de política pública não exigirá necessariamente que se recrute participantes fora da prática normal ou que se implante novos sistemas para prover as intervenções ou monitorar os resultados.

O Behavioural Insights Team tem trabalhado com diversos departamentos governamentais para realizar experiências com pouco dispêndio adicional

do tempo dos membros da equipe. Por exemplo, em experimentos em que a equipe tem de realizar com autoridades locais − HMRC, DVLA e o Serviço de Tribunais (e está para realizar com o Job Centre Plus) −, os dados já são rotineiramente coletados e os processos já estão implantados para apresentar intervenções, quer seja uma carta, uma multa ou um serviço de consultoria para pessoas desempregadas.

Quando se consideram os recursos adicionais que podem ser necessários para realizar um EAC, é preciso lembrar de que eles são geralmente a melhor forma de estabelecer se um programa oferece um bom retorno. Em alguns casos, um experimento pode levar a concluir que um programa é demasiado dispendioso para ser realizado, se os benefícios extras do tratamento forem insignificantes. Em outros, uma experiência pode demonstrar que um programa oferece excelente aplicação de recursos e, por isso, deve ser expandido.

Ao demonstrar quanto mais ou menos eficaz o tratamento foi em relação ao *status quo*, os formuladores de políticas podem determinar se o custo da intervenção justifica os benefícios. Em vez de considerar quanto o EAC custa para ser realizado, pode ser mais apropriado perguntar: *quais são os custos de não fazer um EAC?*[3]

2.2.3 Existem vantagens éticas em usar EACs

Algumas pessoas têm objeção aos EACs nas políticas públicas por considerarem antiético não oferecer um novo tratamento para pessoas que poderiam se beneficiar dele. Este é particularmente o caso quando os recursos adicionais são gastos em programas que poderiam melhorar a saúde, renda ou educação de um grupo.

É legítimo dizer que é questionável não oferecer um tratamento ou intervenção para alguém que acreditamos que possa se beneficiar dele. Este artigo não argumenta que devemos fazer isto quando sabemos que um tratamento já é comprovadamente benéfico.

Entretanto, é preciso ter clareza acerca dos limites de nosso conhecimento e considerar que não se pode ter convicção da eficiência de um tratamento até que este seja exaustivamente testado.

Por vezes, as intervenções consideradas eficazes se mostraram ineficazes ou até mesmo prejudiciais (boxes 6 e 7). Este pode ser o caso, inclusive, de políticas que se supunha, intuitivamente, terem resultado garantido. Por exemplo, têm sido usados incentivos para encorajar alunos adultos a participar de aulas de alfabetização, mas, quando o EAC desta política foi

3. Isso pode ser formalmente estimado ao se comparar o custo de uma experiência, por exemplo, diante de uma estimativa do dinheiro que seria gasto se o tratamento fosse implementado, mas não tivesse benefício.

realizado, constatou-se que os participantes que receberam incentivos participaram de aproximadamente duas aulas a menos por período que o grupo sem incentivo (Brooks *et al.*, 2008).

Nessa experiência, o uso de pequenos incentivos monetários não apenas desperdiçou recursos, como também reduziu a frequência às aulas. Não oferecer o tratamento foi melhor que oferecê-lo, e se um experimento nunca tivesse sido realizado, prejuízos poderiam ter sido causados aos alunos adultos com a melhor das intenções, e sem jamais saber que se estava fazendo isso.

Vale também observar que políticas são geralmente introduzidas aos poucos, de forma gradativa, com algumas áreas começando antes, sem que estas introduções por fases sejam consideradas como antiéticas. A introdução do programa Sure Start é um exemplo disto.

Aliás, uma introdução por fases no contexto de um EAC é mais ética, pelo fato de gerar novas informações de alta qualidade que podem ajudar a demonstrar que um tratamento é eficaz em termos de custo.

2.2.4 Os EACs não precisam ser complicados ou difíceis de realizar

Os EACs, em sua forma mais simples, são de realização bastante objetiva. Entretanto, existem armadilhas que indicam que algum apoio especializado é aconselhável desde o início.

Algumas dessas armadilhas são apresentadas na próxima seção, mas não são maiores que as enfrentadas em qualquer outra forma de avaliação de resultados e podem ser superadas com o apoio correto. Isto pode envolver, por exemplo, um contato com profissionais (a exemplo dos que integram o Behavioural Insights Team) que poderão orientar o desenho do experimento e a comunicação entre os formuladores de políticas e os acadêmicos experientes, realizar EACs e ajudar a orientar o projeto de um experimento. Muito frequentemente, os acadêmicos têm prazer em auxiliar um projeto que lhes proporcionará nova evidência em uma área de interesse para sua pesquisa, ou a possibilidade de um artigo acadêmico publicado.

O esforço inicial para construir uma randomização e definir claramente os resultados antes de um piloto ser iniciado é, geralmente, tempo bem gasto. Se um EAC não é realizado, então qualquer tentativa de avaliar o impacto de um tratamento será difícil, dispendiosa e enviesada – será necessário usar modelos complexos para identificar os efeitos que podem ter causas externas múltiplas. Seria muito mais eficiente investir esforço no desenho de um EAC *antes* de a política ser implementada.

BOX 8
Family Nurse Partnership: construindo uma avaliação rigorosa para um programa mais amplo

A Family Nurse Partnership (FNP) é um programa preventivo para mães de primeira viagem em situação vulnerável. Desenvolvido nos Estados Unidos, envolve visitas domiciliares estruturadas e intensivas, feitas por enfermeiras especialmente treinadas, desde o início da gravidez e até que a criança atinja 2 anos de idade. Diversos EACs nos Estados Unidos[1] têm mostrado benefícios significativos para jovens famílias necessitadas e economia substancial de recursos. Por exemplo, as crianças beneficiadas pela FNP têm melhor desenvolvimento socioemocional e desempenho educacional, além de menor probabilidade de se envolver no crime. Igualmente, as mães têm menos filhos, intervalos maiores entre os nascimentos, mais probabilidade de ser empregadas e menos probabilidade de se envolver no crime.

A FNP é oferecida no Reino Unido desde 2007, geralmente através dos centros Sure Start Children. E o Departamento de Saúde comprometeu-se a ampliar o número de mães jovens que recebem apoio por meio deste programa para até 13 mil (de cada vez) em 2015. Durante o período, o departamento financia uma avaliação de EACs do programa, para definir se a FNP beneficia famílias além dos serviços universais e se oferece benefícios que superam os custos. O experimento envolve dezoito lugares do Reino Unido e aproximadamente 1.650 mulheres, o maior experimento até hoje com a FNP. Conforme informado em 2013, as medidas dos resultados incluem: tabagismo durante a gravidez, amamentação, admissões em hospitais por ferimentos, ingestões acidentais, outras gravidezes e desenvolvimento da criança aos 2 anos.

Elaboração dos autores.
Nota: [1] Para uma síntese das pesquisas das Nações Unidas sobre esse programa, ver MacMillan *et al*. (2009).

3 REALIZAÇÃO DE UM EAC: NOVE PASSOS-CHAVE

3.1 Como conduzir um experimento aleatório controlado?

A seção 2 deste trabalho trata do uso de EACs na política pública. A seção 3 versa sobre como conduzir um EAC; contudo, não pretende ser abrangente. Em vez disso, apresenta os passos necessários pelos quais qualquer EAC deve passar e aponta para as áreas nas quais um formulador de políticas pode desejar buscar orientação mais especializada.

Identificam-se nove passos que qualquer EAC precisará para ser executado. Muitos destes nove passos são bem conhecidos por qualquer um que implante uma avaliação de política bem definida – por exemplo, a necessidade de deixar claro, desde o início, o que a política procura alcançar.

Outros, entretanto, são menos conhecidos; em particular, a necessidade de alocar aleatoriamente o tratamento sendo testado para diferentes grupos. Os passos estão resumidos a seguir e desenvolvidos mais detalhadamente nas subseções seguintes.

- Testar
    1) Identificar duas ou mais intervenções para comparar (por exemplo, política antiga *versus* nova; variações diferentes de uma política).
    2) Determinar o resultado que a política pretende influenciar e como será mensurado no experimento.
    3) Decidir sobre a unidade de randomização: se randomizar a intervenção e os grupos de controle no nível de indivíduos, instituições (por exemplo, escolas) ou áreas geográficas (por exemplo, autoridades locais).
    4) Determinar quantas unidades (pessoas, instituições ou áreas) são necessárias para obter resultados robustos.
    5) Atribuir cada unidade a uma das intervenções, utilizando um método de randomização robusto.
    6) Aplicar as intervenções nos grupos escolhidos.
- Aprender
    7) Mensurar os resultados e determinar o impacto de intervenções.
- Adaptar
    8) Adaptar a política de acordo com suas constatações.
    9) Voltar ao passo 1 para melhorar continuamente seu entendimento das intervenções efetivas.

3.2 Testar

Passo 1: identificar duas ou mais intervenções para comparar

Os EACs são realizados quando há incerteza sobre qual é a melhor de duas ou mais intervenções e envolvem comparar estas intervenções entre si. Geralmente, os experimentos são realizados para comparar um novo tratamento diante da prática corrente. O novo tratamento pode ser uma pequena mudança, ou um conjunto de pequenas mudanças na política usual; ou pode ser uma abordagem nova que tenha sido bem-sucedida em um país ou contexto diferente, ou que tenha apoio teórico sólido.

Antes de desenhar um EAC, é importante considerar o que é atualmente conhecido sobre a eficiência do tratamento que se pretende testar. Pode ser, por exemplo, que os EACs já tenham sido realizados em contextos similares, e a intervenção tenha se mostrado eficaz ou ineficaz. A pesquisa existente pode também ajudar a desenvolver a própria intervenção. Um bom ponto

de partida são os arquivos da Campbell Collaboration,[4] que apoiam os formuladores e que aplicam as políticas ao resumir a evidência existente sobre políticas sociais.

É também importante que os experimentos sejam realizados na forma que seria aplicada caso o experimento fosse bem-sucedido. Geralmente, existe uma tentação de realizar o EAC usando uma política ideal, perfeita, que é tão dispendiosa que nunca poderia ser realizada em nível nacional. Mesmo que os recursos estivessem disponíveis, este EAC não seria informativo, porquanto os resultados não seriam aplicáveis ao mundo real.

É preciso se certificar de que os resultados do experimento refletirão o que pode ser alcançado caso a política seja considerada eficaz e, então, realizada mais amplamente. Para que os resultados sejam generalizáveis e relevantes para todo o país, o tratamento deve ser representativo, assim como o entusiasmo com o qual os profissionais a aplicam e a forma como os dados são coletados.

O Behavioural Insights Team, ao conduzir um EAC de política pública, gasta certo período de tempo junto às organizações na linha de frente, tanto para entender o que é provavelmente viável, como para aprender com a equipe, que pode ter desenvolvido novos métodos potencialmente eficazes, mas ainda não testados, para alcançar os resultados desejáveis da política pública.

BOX 9
Comparação das diferentes opções políticas e teste de pequenas variações em uma política

Um EAC não é necessariamente um teste entre fazer algo e não fazer nada. Muitas intervenções podem ser melhor que nada. Em vez disso, os experimentos podem ser usados para estabelecer quais opções entre algumas intervenções são as melhores.

Em alguns casos, pode-se ter interesse em responder a grandes perguntas sobre qual política é a mais apropriada. Por exemplo, suponha-se que houvesse os recursos para aprimorar os laboratórios de informática de todas as escolas secundárias, ou contratar mais professores, mas não ambos. Seria possível realizar uma experiência de três alternativas (figura 3), com um grupo de controle (algumas escolas continuando com os computadores atuais e o mesmo número de professores) e dois grupos de intervenção (escolas que receberam uma melhoria do laboratório ou mais professores). Isto nos possibilitaria determinar tanto se a nova política foi eficaz quanto se ofereceu a melhor aplicação dos recursos.

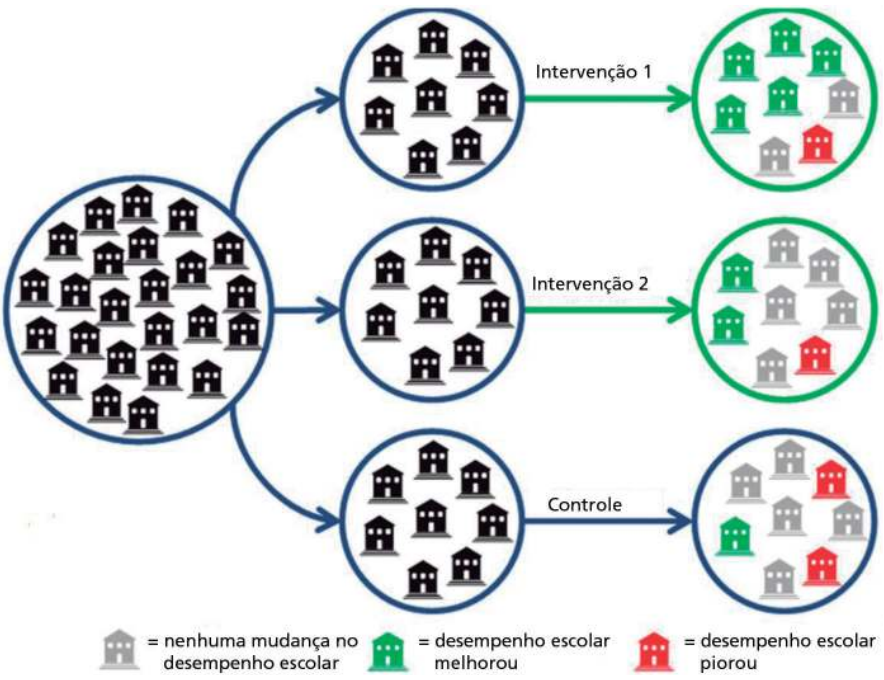
As oportunidades para a sintonia fina de políticas geralmente surgem quando se está prestes a introduzir mudanças – é uma hora ideal para testar variações menores para assegurar que as mudanças que finalmente introduziremos terão o melhor efeito.

(Continua)

4. Disponível em: <http://www.campbellcollaboration.org/library.php>.

(Continuação)

FIGURA 3
O projeto de um EAC testa se a melhoria nos laboratórios de informática das escolas (intervenção 1) ou o emprego de mais professores (intervenção 2) melhora o desempenho acadêmico



Elaboração dos autores.

Passo 2: definir o resultado que a política pretende influenciar e como será mensurado no experimento

Em qualquer experimento, é essencial que se defina de antemão exatamente quais resultados se buscam e como mensurá-los. Por exemplo, no contexto de política educacional, a mensuração de resultado pode se basear em resultados de exames padronizados. Para políticas relacionadas à eficiência de energia doméstica, a mensuração de resultado pode ser o consumo de energia residencial.

É importante ser claro sobre como e quando os resultados serão mensurados ainda no estágio de desenho do experimento, além de manter estes critérios predeterminados no estágio de análise. É também importante assegurar que a forma como os resultados serão mensurados para todos os grupos seja exatamente a mesma – tanto em termos do processo de mensuração como nos padrões seguidos.

A predefinição das medidas de resultado não tem apenas sentido prático. Existem também bons motivos científicos que a tornam crucial para o sucesso

de um EAC. Com o transcorrer do tempo, sempre haverá flutuações aleatórias em dados coletados rotineiramente. Ao final do experimento, há muitos dados sobre várias coisas diferentes, e, quando há tantos dados, é inevitável que alguns números melhorem – ou piorem – simplesmente pela variação aleatória no transcorrer do tempo.

Sempre que essa variação aleatória ocorrer, pode ser tentador pegar alguns números que melhoraram, simplesmente ao acaso, e considerá-los como evidência de sucesso. Entretanto, fazer isto viola os pressupostos dos testes estatísticos, porque aumenta as chances de encontrar um resultado positivo. A tentação de sobreinterpretar os dados e atribuir significado à variação aleatória é evitada ao predeterminar a medida de resultados. Os testes estatísticos podem ser, então, usados para analisar quanto da variação é simplesmente devido ao acaso.

BOX 10
Tirando vantagem das oportunidades naturais para EACs

Por vezes, os limites na aplicação da política proporcionam o contexto ideal para um experimento. Por exemplo, as restrições financeiras e/ou os aspectos práticos podem fazer com que uma implantação escalonada seja a opção preferida. Se houver facilidade de monitorar os resultados em todas as áreas que receberão a intervenção, e for possível aleatorizar qual será a primeira a recebê-la, uma implantação escalonada pode ser explorada para realizar um desenho de pesquisa escalonado.

Por exemplo, o serviço de liberdade condicional na área de Durham queria testar uma nova abordagem. As restrições de recursos impediram todos os seis centros de condicional de receber a nova orientação e treinamento ao mesmo tempo. A abordagem mais apropriada, e cientificamente a mais robusta, foi atribuir aleatoriamente aos seis centros uma posição em uma lista de espera. Todos os centros, finalmente, receberam o treinamento, mas como, em vez da conveniência administrativa, a alocação aleatória determinou quando cada centro recebeu o treinamento, pôde ser realizada uma avaliação robusta dos efeitos do novo serviço na taxa de reincidência.[1]

Elaboração dos autores.
Nota: [1] Resultados finais ainda serão publicados. Para detalhes do desenho do estudo, ver Pearson *et al*. (2010).

Quando da decisão da medida de resultado, é também importante identificar um resultado que realmente seja esperado, ou o mais próximo que se consegue chegar, em vez de uma mensuração de processos intermediários. Por exemplo, em um experimento para verificar se oficiais dos serviços de liberdade condicional trabalhando em conjunto no atendimento para alcoólatras poderiam reduzir a reincidência, é possível mensurar itens como: encaminhamento para serviço de atendimento a alcoólatras, participações nos serviços de atendimento a alcoólatras, ingestão de álcool, ou reincidência.

Nesse caso, a reincidência é o resultado que consiste no centro da preocupação, mas os dados podem ser mais difíceis de serem coletados, e qualquer benefício pode levar anos para se tornar aparente. Por este motivo, pode-se considerar a mensuração da participação em serviços de atendimento a alcoólatras como uma *proxy* para o resultado de interesse. Alternativamente, pode-se mensurar ambos: a participação no serviço, como resultado provisório; e, então, resultados de acompanhamento de longo prazo para comportamento reincidente 24 meses após. *O número de encaminhamentos para o serviço de atendimento a alcoólatras* pode ser a coisa mais fácil de mensurar, mas, apesar de imediato, não seria muito informativo se a reincidência for o que realmente importa (box 11).

A questão de qual medida de resultado usar geralmente se beneficia dos debates entre os acadêmicos (que sabem, tecnicamente, o que pode funcionar melhor em um experimento) e os formuladores de políticas (que sabem que tipos de dados estão disponíveis, e quanto pode custar para coletar).

BOX 11
A favor (e contra) o uso de *proxies*

Uma *proxy* é aquela que substitui a medida de interesse verdadeira – por exemplo, os índices de recondenação são usados como substitutos para índices de reincidência –, porque são bem mais fáceis de mensurar (já que as pessoas podem nunca ser pegas pelos crimes que cometerem). O argumento para usar uma *proxy* é mais robusto quando houver boa evidência de que se trata de um forte indicador do resultado final de interesse. Infelizmente, usar as medidas autoinformadas de mudança comportamental, apesar de serem fáceis de ser mensuradas, pode ser um índice pobre da mudança real do comportamento. Devido ao viés de "desejabilidade social", as pessoas podem ser motivadas a informar exageradamente, por exemplo, a quantidade de exercício que fazem, após terem participado de um programa de condicionamento físico.

Se as *proxies* forem necessárias porque os resultados finais são de longo prazo, sempre vale a pena acompanhar estes resultados de longo prazo para verificar os resultados intermediários. Existem diversos casos na medicina em que as experiências iniciais utilizando variáveis *proxy* foram enganosas. Por exemplo, oferecer aos pacientes com osteoporose tratamento com fluoreto foi considerado eficaz, pois aumentava a densidade óssea. Como um dos indicadores clínicos principais da osteoporose, a densidade óssea foi considerada uma *proxy*. Entretanto, tem sido demonstrado que o tratamento com fluoreto, na realidade, leva a um aumento em alguns tipos de fraturas; o resultado final que pacientes osteoporóticos buscam evitar (Riggs, Hodgson e O'Fallon, 1990; Rothwell, 2005).

Elaboração dos autores.

Passo 3: decidir sobre a unidade de randomização

Após decidir que resultado será mensurado (passo 2), é o momento de decidir quem ou o que o que será randomizado. Isto é conhecido como a unidade de randomização.

A unidade de randomização é geralmente individual; por exemplo, quando os indivíduos são escolhidos aleatoriamente para receber um ou dois tratamentos médicos, ou um ou dois programas educacionais. Entretanto, a unidade de randomização pode ser também um grupo de pessoas em uma instituição, especialmente se o tratamento é algo que é mais bem aplicado a um grupo. Por exemplo, escolas inteiras podem ser escolhidas aleatoriamente para receber um novo método de ensino, ou o atual; *job centres*[5] inteiros podem ser escolhidos aleatoriamente para oferecer um novo programa de treinamento, ou o atual. Por último, a unidade de randomização pode ser toda a área geográfica: por exemplo, as autoridades locais podem ser escolhidas aleatoriamente para realizar um ou mais programas novos de saúde ou métodos diferentes de reciclagem de lixo (box 12).

BOX 12
Explorando variações locais da política

As autoridades locais estão bem posicionadas para testar novas políticas na área. Ao colaborarem com outras autoridades locais para experimentar políticas diferentes ou escolher aleatoriamente diferentes ruas ou regiões para diferentes intervenções, as autoridades locais podem usar a metodologia de EACs para determinar quais políticas são eficazes.

Um exemplo dessa abordagem é a experiência realizada pelo governo municipal de North Trafford para comparar métodos diferentes de promover a reciclagem de lixo. A unidade randomizada nesta experiência foi *ruas inteiras*. Metade das ruas foram escolhidas aleatoriamente para receber uma campanha de incentivo à reciclagem de lixo. Os índices de reciclagem foram mais elevados neste grupo, comparado às quase 3 mil residências que não receberam a campanha.

O aumento a curto prazo foi de 5%, e os parceiros acadêmicos julgaram que a campanha de coleta custou em torno de £ 24,00 por cada residência adicional que começou a reciclar (Cotterill, John e Liul, 2008; John *et al*., 2011). Com base nestas informações, o governo local torna-se, então, capaz de determinar se a redução de custos de aterro sanitário associados à campanha de coleta justificaria os custos de oferecê-la mais amplamente.

Elaboração dos autores.

Ao final do experimento, os resultados podem ser mensurados em indivíduos ou para toda a unidade de randomização, dependendo do que for mais prático e mais preciso. Por exemplo, embora turmas inteiras possam ser aleatoriamente escolhidas para receber diferentes métodos de ensino, os resultados da aprendizagem dos alunos podem ser avaliados quando se calculam os resultados, para maior exatidão.

A questão da escolha da unidade de randomização dependerá de considerações práticas. Em experiências clínicas, por exemplo, é geralmente possível dar

5. Nota dos revisores técnicos: *job centres* combinam os serviços de uma agência de empregos pública com os de pagamento de seguro-desemprego e outros benefícios.

a indivíduos diferentes o placebo ou o medicamento que estiver sendo testado. Mas em experiências de política pública, nem sempre pode ser possível fazer isso. A seguir, são analisados dois exemplos de formas diferentes nas quais o Behavioural Insights Team decidiu sobre qual unidade de randomização usar.

1) *Individual*: ao analisar mensagens diferentes nas cartas de cobrança de impostos, é obviamente possível enviar cartas diferentes para diferentes indivíduos, e, por isso, a unidade de randomização consistia em devedores individuais.

2) *Instituição*: em um experimento para apoiar as pessoas a conseguirem emprego em *job centres*, não é possível escolher aleatoriamente intervenções diferentes para diferentes candidatos a emprego, e, por isso, a unidade de randomização será as equipes do centro de emprego (ou seja, as equipes dos consultores que ajudam os candidatos na busca de emprego).

Como nos outros passos, seria útil discutir a unidade de randomização com um pesquisador acadêmico. Seria também importante analisar como a decisão de escolher uma determinada unidade interage com outras considerações. Mais importante, esta decisão afetará quantas pessoas precisarão se envolver na experiência: ter instituições ou áreas como sua unidade de estudo quase sempre significará que é necessária uma amostra maior de indivíduos, e também são necessários métodos especiais de análise.

Há, ainda, outras considerações que podem ser feitas; por exemplo, em uma avaliação de incentivos à frequência em cursos voltados à educação de adultos, os pesquisadores escolhem randomizar todas as turmas, apesar de ser possível randomizar participantes individuais. Esta decisão foi tomada para evitar que aqueles no grupo sem incentivos se ressentissem por outros alunos na turma estarem recebendo um incentivo, e eles não. Obviamente, isto poderia afetar negativamente o índice de frequência nas classes, e poderia se observar um efeito em virtude deste problema, e não devido ao incentivo.

BOX 13
Quando a unidade de randomização deve ser grupos em lugar de indivíduos

Vermes como a tênia infectam quase um quarto da população mundial, a maioria em países em desenvolvimento. É uma causa comum do afastamento escolar. E os pesquisadores dos Estados Unidos colaboraram com o Ministério da Saúde daquele país para determinar se oferecer tratamento para eliminar vermes nas crianças aumentaria a frequência escolar.

(Continua)

Vermes como a tênia infectam quase um quarto da população mundial, a maioria em países em desenvolvimento. É uma causa comum do afastamento escolar. E os pesquisadores dos Estados Unidos colaboraram com o Ministério da Saúde daquele país para determinar se oferecer tratamento para eliminar vermes nas crianças aumentaria a frequência escolar.

Foi realizado um EAC no qual todas as escolas receberam tratamento maciço de eliminação de vermes ou continuaram como estavam. Neste caso, a randomização individual teria sido inapropriada – se alguns alunos tivessem os vermes eliminados e outros não, a probabilidade de que os participantes do controle fossem contaminados pode ter sido artificialmente reduzida pelo fato de seus colegas estarem sem vermes.

Setenta e cinco escolas primárias na área rural do Quênia participaram do estudo, que demonstrou que o programa de eliminação de vermes reduziu o número de faltas em um quarto (Miguel e Kremer, 2004). Os aumentos na frequência escolar foram particularmente marcantes nas crianças mais jovens. Este estudo demonstrou que um ano adicional de frequência escolar poderia ser alcançado graças à eliminação de vermes a um custo de US$3,50 por aluno, representando um método altamente eficaz em termos de custo para aumentar a frequência escolar, enquanto outros programas, como uniformes escolares gratuitos, custam mais de US$100,00 por aluno para gerar efeitos similares (Karlan e Appel, 2011).

Elaboração dos autores.

Além disso, é crucial que os indivíduos sejam recrutados para o estudo antes de ser feita a randomização, ou o experimento deixa de ser robusto.

Por exemplo, se as pessoas que realizam uma experiência souberem a qual grupo um possível participante seria alocado, antes deste participante ser recrutado para o estudo, isto pode afetar a decisão de recrutá-los. Um pesquisador ou membro do pessoal de campo que acreditar veementemente no novo tratamento pode escolher – talvez inconscientemente –não recrutar participantes que ele acredita serem "sem esperança" no novo grupo de intervenção. Isto significaria que os participantes em cada grupo "aleatório" não são mais representativos. Este tipo de problema pode ser evitado garantindo-se, simplesmente, que os participantes sejam recrutados primeiramente para a experiência, e só então randomizados.

Passo 4: determinar quantas unidades são necessárias para obter resultados robustos

Para se tirar conclusões de um EAC, a experiência deve ser realizada com um tamanho de amostra suficiente. Se o tamanho da amostra for suficientemente grande, é improvável que o efeito do tratamento tenha sido resultado do acaso.

Caso se decida que a unidade de randomização será instituições ou áreas, é muito provável que se precise de um número maior de pessoas no experimento que na opção de se randomizar por indivíduo. Simples cálculos de potência estatística preliminares ajudarão a determinar quantas unidades (indivíduos, instituições etc.)

devem ser incluídas no tratamento e nos grupos de controle. Recomendamos trabalhar com acadêmicos que tenham experiência em EACs para garantir que este cálculo técnico seja feito corretamente.

Se a política produzir um grande benefício (um grande "efeito substantivo"), será possível detectar isto usando um experimento com tamanho de amostra relativamente pequeno. Detectar diferenças mais sutis (efeitos substantivos pequenos) entre as intervenções exigirá números maiores de participantes; por isto, é importante, desde o início, não ser demasiado otimista sobre o provável sucesso do tratamento. Muitas intervenções – se não a maioria – têm efeitos relativamente pequenos.

Como exemplo de quantos participantes são necessários para uma experiência, suponha-se a alocação aleatória de oitocentas pessoas em dois grupos de quatrocentas pessoas em cada um, isto corresponde a cerca de oito entre dez chances de ver uma diferença de 10%, caso esta diferença existisse.

Por exemplo, imagine-se que o governo quer incentivar as pessoas a votar e quer testar a eficácia de enviar mensagens de texto para eleitores cadastrados, para lembrá-los na manhã de uma eleição. Eles escolhem oitocentos eleitores para observar: quatrocentos no grupo de controle, que não receberão nenhum lembrete extra, e quatrocentos no grupo de tratamento, que receberão mensagens de texto.

Se o comparecimento for de 50% no grupo de controle, com uma amostra desse tamanho, haveria uma chance de 80% de ver uma mudança de 50% a 60% (uma mudança de 10 pontos percentuais). Para detectar uma diferença menor, seria preciso dispor de tamanhos maiores de amostra.

Deve-se considerar quanto custa recrutar cada pessoa adicional e o impacto (tamanho do efeito e economia potencial de custo) do tratamento que está sendo mensurado. Por vezes, detectar mesmo uma diferença modesta é muito útil, particularmente se o próprio tratamento custar pouco ou nada. Por exemplo, ao mudar o estilo ou conteúdo de uma carta para incentivar o pagamento imediato de impostos, o custo adicional é muito pequeno, uma vez que os custos com postagem são os mesmos e já se está automaticamente coletando os dados do resultado (neste caso, as datas de pagamento). Em contrapartida, oferecer uma orientação individualizada para pessoas que estão recebendo seguro-desemprego com o objetivo de aumentar a proporção daquelas que conseguem um trabalho em tempo integral é relativamente dispendioso, e, além disto, espera-se um efeito imensamente maior para valer a pena realizar uma experiência. Entretanto, mesmo para intervenções dispendiosas, se os impactos hipotéticos são pequenos em termos de tamanho do efeito, mas potencialmente grandes em termos de economia (por exemplo, reduções no número de pessoas requerendo benefícios), pode haver um motivo robusto para realizar um EAC.

Passo 5: atribuir cada unidade a uma das intervenções políticas, usando um método robusto de randomização

A alocação aleatória das unidades de tratamento e controle é o passo-chave que torna o EAC superior a outros tipos de avaliação de política: oferece maior confiança de que o grupo de intervenção política seja equivalente, com respeito a todos os fatores-chave. No contexto de política educacional, por exemplo, isto pode incluir situação socioeconômica, gênero e experiência.

O viés pode surgir sob diversas formas durante o processo de randomização; assim, para evitar problemas posteriores, é importante certificar-se de que este passo seja feito corretamente desde o início.

Existem evidências de que as pessoas que possuem interesse em um estudo possam procurar alocar as pessoas de forma não aleatória, até inconscientemente. Por exemplo, suponha-se um experimento que esteja alocando pessoas a um tratamento "de volta ao trabalho" com base nos seus números de seguridade nacional, e números ímpares forem obter o novo tratamento. Devido ao desejo de fazer o novo tratamento parecer bom, pode acontecer de a pessoa que recrutar o participante, consciente ou inconscientemente, excluir do experimento certas pessoas com um número ímpar, caso suspeite que estas não se sairão bem.

Isso introduziria um viés no experimento, e, assim, o método de randomização deve ser resistente a esta interferência. Existem muitas organizações independentes, como unidades clínicas de experimentos, que podem ajudar a estabelecer um método seguro de randomização para evitar este problema de "ocultação de alocação fraca". Normalmente, isto envolverá um gerador de número aleatório, que determina em que grupo um participante será alocado, e somente após ter sido recrutado para a experiência (pelos motivos descritos).

Por ocasião da randomização, caso se julgue importante, os passos podem ser realizados para assegurar que os grupos sejam igualmente equilibrados com relação a diferentes características – por exemplo, para se certificar de que exista praticamente a mesma distribuição de idade e sexo em cada grupo. Isto é particularmente importante em experimentos de pequena escala, já que eles têm menos poder estatístico.

BOX 14
Elaborando variações para possibilitar os testes

Os testes envolvem a comparação do efeito de um tratamento (por exemplo, possível nova política) diante de outro (política atual). Um teste sólido, obviamente, exige que variações na política (novas e atuais) possam ser simultaneamente executadas. Em alguns casos, isto é bem simples – algumas escolas podem continuar servindo as atuais merendas escolares, enquanto outras aderem aos novos padrões nutricionais, e o efeito sobre o comportamento das turmas pode ser mensurado. Em outros casos, os sistemas instalados podem tornar difícil oferecer diferentes variações das políticas ao mesmo tempo.

(Continua)

(Continuação)

Por exemplo, embora uma autoridade local possa querer testar a eficácia de simplificar um formulário, seus sistemas de impressão podem ser terceirizados e/ou incapazes de imprimir mais de um modelo. Por este motivo, sugere-se fortemente que, quando desenvolverem sistemas ou procurarem novos contratos de provedores de serviço, os formuladores de política se assegurem de que serão capazes de produzir variações de políticas no futuro. Embora isto possa chegar a um custo inicial desprezível, a capacidade de testar diferentes versões de políticas no futuro provavelmente é mais que justificada. Com isto em mente, a legislação de DWP permite que os sistemas de informática que controlam o Universal Credit (Christensen, 2003) incluam o recurso de fornecer variações, para garantir que o departamento seja capaz de testes para descobrir o que funciona e adaptar seus serviços para refletir isto.

Elaboração dos autores.

Passo 6: introduzir as intervenções políticas nos grupos escolhidos

Uma vez que os indivíduos, as instituições ou as áreas geográficas tenham sido alocados aleatoriamente a um grupo de tratamento ou a um controle, então, é o momento de introduzir o tratamento.

Isso pode envolver, por exemplo, a introdução de novo tipo de política de educação em um grupo de escolas, e não fazer as mudanças em outro lugar. Quando o Behavioural Insights Team testou se as mensagens de texto poderiam melhorar a propensão das pessoas de pagar suas multas de tribunal, por exemplo, os indivíduos nos grupos de intervenção receberam um dos diversos tipos diferentes de mensagem de texto, enquanto os do grupo de controle não receberam texto nenhum.

Uma consideração importante nesse estágio é ter um sistema implantado para monitorar o tratamento, para assegurar que esteja sendo introduzido na forma na qual foi originalmente planejado. No exemplo da mensagem de texto, por exemplo, assegurou-se que os textos corretos estavam indo para as pessoas certas. O uso de uma avaliação de processo para monitorar que o tratamento seja introduzido conforme pretendido garantirá que os resultados sejam o mais significativos possível e os contratempos, corrigidos.

Como nos outros passos, entretanto, é importante garantir que o experimento reproduza a política que será introduzida quando e se esta for ampliada. Por exemplo, na experiência da mensagem de texto, ocorreu que nem sempre se tem os números corretos do telefone celular de todos nos grupos.

Seria tentador gastar tempo e dinheiro adicionais para verificar e procurar esses números faltantes de telefone, mas não teria refletido como o tratamento seria aplicado caso o experimento fosse ampliado, e teria, por conseguinte, feito os resultados parecerem mais bem-sucedidos que seriam na "vida real".

3.3 Aprender

Passo 7: mensurar os resultados e determinar o impacto das intervenções políticas

Uma vez que o tratamento tenha sido introduzido, é preciso mensurar os resultados. O momento e o método de avaliação do resultado devem ter sido decididos antes da randomização. Dependerá de com que rapidez se supõe que o tratamento funcionará, o que varia de acordo com cada tratamento.

Uma experiência de cartas diferentes para pessoas incentivando-as a pagar suas multas pode precisar apenas de acompanhamento de algumas semanas, enquanto a intervenção no currículo pode necessitar de um período escolar ou até mesmo diversos anos.

Além do resultado principal, pode ser útil coletar indicadores de processo. Por exemplo, em um estudo de diferentes serviços de condicional, pode-se coletar dados sobre encaminhamentos para diferentes agências, para ajudar a explicar os resultados. Neste caso, uma redução na reincidência pode ser acompanhada por um aumento correspondente nos encaminhamentos para aulas de controle da agressividade, que podem explicar os resultados. Estes resultados secundários não podem ser interpretados com a mesma certeza que os resultados principais da experiência, mas podem ser usados para desenvolver novas hipóteses para outras experiências (box 15).

Além disso, muitas experiências também envolvem a coleta de dados qualitativos para ajudar a explicar os resultados, apoiar a implementação futura e atuar como guia para outra pesquisa ou melhorar o tratamento. Isto não é necessário, mas se a pesquisa qualitativa for de qualquer forma planejada, é ideal fazê-lo com os mesmos participantes dessas experiências, uma vez haverá mais informações disponíveis.

BOX 15
Uso mais inteligente dos dados

Normalmente, há interesse em verificar se uma política é amplamente eficaz para uma amostra representativa na população em geral. Em alguns casos, entretanto, o objetivo pode ser descobrir se alguns grupos (por exemplo, homens e mulheres, jovens e idosos) reagem de forma diferente de outros. É importante, no início, decidir se há interesse em segmentar a amostra desta forma – fazendo-se isto após os dados terem sido coletados, corre-se um alto risco de que a análise do subgrupo careça de poder estatístico ou validade. Entretanto, caso uma tendência de subgrupo surja inesperadamente (como parecer que os homens reagem mais que as mulheres a lembretes por mensagens de texto para participarem de consultas médicas), pode-se considerar a realização de uma experiência no futuro para descobrir se este é um resultado robusto. Geralmente, vale a pena coletar dados adicionais (por exemplo, idade e gênero), que ajudarão a segmentar a amostra e guiar a pesquisa futura.

(Continua)

(Continuação)

Por vezes, uma tendência antecipada pode surgir dos dados do experimento. Por exemplo, pode acontecer de notarem-se grandes flutuações no tempo na eficácia de um incentivo para melhorar o isolamento térmico das residências, e descobrir que isto está relacionado a variações de temperatura. Esta tendência pode indicar que as pessoas sejam mais receptivas a mensagens sobre isolamento residencial quando o tempo estiver frio. Como se trata de uma análise não planejada, os resultados não podem ser considerados definitivos; entretanto, nenhuma informação deve ser desperdiçada, e este resultado pode ser valioso para a pesquisa futura.

Elaboração dos autores.

3.4 Adaptar

Passo 8: adaptar sua política para refletir suas descobertas

Implementar resultados positivos das intervenções é geralmente mais fácil que convencer as pessoas a suspender políticas que demonstraram ser ineficazes. Qualquer experiência realizada, concluída e analisada deve ser considerada bem-sucedida. Um EAC que não demonstre nenhum efeito ou mesmo demonstre um efeito prejudicial da nova política é tão valioso como um que demonstre um benefício.

A experiência do DWP de apoiar as pessoas que estavam recebendo auxílio-doença foi um estudo "nulo", já que não demonstrou eficácia (box 1). Entretanto, ao considerá-lo um teste apropriado para saber se o tratamento funciona (o que deve ser estabelecido antes da experiência começar), e entendendo-se que o tamanho da amostra foi suficientemente grande para detectar qualquer benefício de interesse (isto, novamente, deve ser estabelecido antes do início), informações úteis certamente são obtidas com este experimento.

Quando as intervenções forem ineficazes, então, o "desinvestimento racional" deve ser considerado e os recursos economizados podem ser gastos de outra forma, em intervenções eficazes. Ademais, esses resultados também funcionam como catalisadores para descobrir outras intervenções eficazes; por exemplo, outras intervenções para ajudar as pessoas com auxílio-doença.

Quando um EAC for concluído, é aconselhável publicar os resultados, com informações completas sobre os métodos do experimento, de forma que os outros possam avaliar se foi um "teste apropriado" do tratamento. É também importante incluir uma descrição completa do tratamento e dos participantes, de forma que outros possam implementar o programa com confiança em outras áreas, se assim desejarem.

Um documento útil que pode orientar a redação do relatório do EAC é o *CONSORT Statement*,[6] que é usado em experiências médicas e também, de forma

6. Disponível no *site* do Consolidated Standards of Reporting Trials (CONSORT), em: <http://www.consort-statement.org/consort-statement>.

crescente, em experiências não médicas. Seguir a orientação do Consolidated Standards of Reporting Trials (CONSORT) assegurará que as partes-chave da experiência e as intervenções sejam descritas de forma suficientemente precisa, para permitir a reprodução do experimento ou a implementação do tratamento em outra área.

De forma ideal, o protocolo da experiência deve ser publicado antes de a experiência começar, para que as pessoas possam apresentar críticas ou aperfeiçoamentos. A publicação do protocolo também torna claro que o resultado principal informado foi o escolhido antes de a experiência começar.

Passo 9: retornar ao passo 1 para melhorar continuamente seu entendimento do que funciona

Em vez de ver o EAC como uma ferramenta para avaliar um único programa em um determinado ponto no tempo, é útil pensar nele como parte de um processo contínuo de inovação e aperfeiçoamento. A replicação dos resultados de um experimento é particularmente importante se o tratamento for oferecido a um segmento da população diferente daquele que estava envolvido no EAC original. É também útil se basear nos experimentos anteriores para identificar novas formas de melhorar os resultados pertinentes, quando os EACs forem usados para identificar que aspectos de uma política têm maior impacto. Em trabalho recente com HMRC, por exemplo, o Behavioural Insights Team procurou entender que mensagens são mais eficazes para ajudar as pessoas a cumprir suas obrigações tributárias.

Diversas lições foram aprendidas sobre o que funciona melhor – por exemplo, formulários e cartas o mais simples possível e informar os devedores que a maioria dos outros em sua vizinhança já pagou seus impostos.

Entretanto, em vez de contar com essas lições e assumir que a perfeição foi alcançada, é mais útil pensar no potencial para maior aperfeiçoamento. Por exemplo, existem outras formas de simplificar os formulários e tornar mais fácil para os contribuintes pagarem seus compromissos, ou existem outras mensagens mais eficazes com tipos diferentes de contribuintes?

O mesmo tipo de pensamento pode ser aplicado a todas as áreas políticas – desde melhorar o resultado de testes padronizados até ajudar as pessoas a encontrarem (e se manterem nos) empregos.

Melhoria contínua, nesse sentido, é o aspecto final, e talvez o mais importante, da metodologia "testar, aprender e adaptar", porque supõe que nunca se sabe tanto quanto seria possível na área das políticas públicas.

BOX 16
Reduzir a mortalidade de pacientes em asilos

As vacinas contra gripe são rotineiramente oferecidas a grupos de risco, inclusive os idosos, quando o inverno se aproxima. Nos asilos, entretanto, o vírus da gripe pode ser introduzido através dos funcionários. Em 2003, foi realizado um EAC para determinar se o custo de levar equipes para vacinar os funcionários: *i*) aumentaria os índices de vacinação destes; e *ii*) teria efeitos positivos na saúde dos idosos. Mais de quarenta asilos foram alocados aleatoriamente tanto para continuarem da forma usual (sem uma iniciativa de vacinação dos funcionários), como para implantar uma campanha para conscientizar os funcionários sobre as vacinas para gripe e oferecer consultas para inoculação. Após duas temporadas de gripe, a vacinação de funcionários foi significativamente maior em asilos que instituíram a campanha para vacinação de gripe, talvez sem surpresa. O mais importante é que a mortalidade dos residentes resultou também mais baixa, com cinco mortes a menos para cada cem residentes (Hayward *et al*., 2006). Esta pesquisa contribuiu para uma recomendação nacional de vacinar os funcionários de asilos e é citada, internacionalmente, como parte da justificativa de recomendações contínuas para vacinar funcionários na área de assistência à saúde.

Elaboração dos autores.

REFERÊNCIAS

BANERJEE, A.; DUFLO, E. **Poor economics**: a radical rethinking of the way to fight global poverty. New York: Public Affairs, 2011.

BANERJEE, A. V. *et al*. Remedying education: evidence from two randomised experiments in India. **Quarterly journal of economics**, v. 122, n. 3, p. 1.235-1.264, 2007.

BROOKS, G. *et al*. Randomised controlled trial of incentives to improve attendance at adult literacy classes. **Oxford review of education**, v. 34, n. 5, p. 493-504, 2008.

CHRISTENSEN, C. **The innovator's dilemma**: the revolutionary book that will change the way you do business. New York: Harper Business, 2003.

COTTERILL, S.; JOHN, P.; LIULN, H. **How to get those recycling boxes out**: a randomised controlled trial of a door to door recycling campaign. *In*: RANDOMISED CONTROLLED TRIALS IN THE SOCIAL SCIENCES: METHODS AND SYNTHESIS. York, 30 Sept-2 Oct. 2008.

DAVENPORT, T. H.; HARRIS, J. G. **Competing on analytics**: the new science of winning. Allston: HBS, 2007.

DELTA AIRLINES. **Delta Airlines magazine**, n. 915, p. 22, 2007.

DWP – DEPARTMENT FOR WORK AND PENSIONS. **Impacts of the job retention and rehabilitation pilot**. London: DWP, 2006a. (Research Report, n. 342).

_____. **Jobseekers allowance intervention pilots quantitative evaluation**. London: DWP, 2006b. (Research Report, n. 382).

EDWARDS, P. *et al*. Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury: outcomes at 6 months. **Lancet**, v. 365, p. 1.957-1.959, 2005.

FINCKENAUER J. O. **Scared Straight and the Panacea Phenomenon**. Englewood Cliffs: Prentice-Hall, 1982.

HARFORD, T. **Adapt**: why success always starts with failure. London: Little, 2011.

HAYWARD, A. *et al*. Effectiveness of influenza vaccine programme for care home staff to prevent death, morbidity and health service use among residents; cluster randomised control trial. **British medical journal**, v. 333, n. 7.581, p. 1.241-1.247, 2006.

JOHN, P. *et al*. **Nudge, nudge, think, think**: using experiments to change civic behaviour. London: Bloomsbury Academic, 2011.

KARLAN, D.; APPEL, J. **More than good intentions**: how a new economics is helping to solve global poverty. New York: Dutton, 2011.

LUCA, M. **Reviews, reputation, and revenue**: the case of Yelp.com. Allston: HBS, 2011. (HBS Working Paper, n. 12-016). Disponível em: <http://www.hbs.edu/faculty/Publication%20Files/12-016.pdf>.

MACMILLAN, H. L. *et al*. Interventions to prevent child maltreatment and associated impairment. **Lancet**, v. 363, n. 9.659, p. 250-266, 2009.

MIGUEL, E.; KREMER, M. Worms: identifying impacts on education and health in the presence of treatment externalities. **Econometrica**, n. 72, p. 159-217, 2004. Disponível em: <http://goo.gl/bIRUfd>.

PEARSON, D. *et al*. A parable of two agencies, one of which randomises. **Annals of the American Academy of Political & Social Sciences**, n. 628, p.11-29, 2010.

PETROSINO, A.; TURPIN-PETROSINO, C.; BUEHLER, J. **Scared Straight and other juvenile awareness programs for preventing juvenile delinquency**. Campbell Review Update I. The Campbell Collaboration Reviews of Intervention and Policy Evaluations (C2-RIPE). Philadelphia: Campbell Collaboration, 2003.

RIGGS, B. L.; HODGSON, S. F.; O'FALLON, W. M. Effect of fluoride treatment on fracture rate in postmenopausal women with osteoporosis. **New England journal of medicine**, n. 322, p. 802-809, 1990.

ROTHWELL, P. M. External validity of randomised controlled trials: "to whom do the results of this trial apply?" **Lancet**, n. 365, p. 82-93, 2005.

SHEPHERD, J. The production and management of evidence for public service reform. **Evidence and policy**, v. 3, n. 2, p. 231-251, 2007.

TALEB, N. N. **The black swan**: the impact of the highly improbable. London: Allen Lane, 2007

THE SOCIAL RESEARCH UNIT. **Youth justice**: cost and benefits. Investing in children, 2.1. Dartington: The Social Research Unit, Apr. 2012. Disponível em: <http://www.dartington.org.uk/investinginchildren>.

Texto especial. Selecionado pelos editores em março de 2013.