https://doi.org/10.18222/eae.v34.9956

DELINEAMENTOS EXPERIMENTAIS NA AVALIAÇÃO DE POLÍTICAS PÚBLICAS: USOS E ABUSOS[1]

PAULO DE MARTINO JANNUZZI[I]

[I] Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro-RJ, Brasil; *paulo.jannuzzi.br@gmail.com*

RESUMO

O objetivo deste ensaio é discutir as potencialidades e, sobretudo, os limites do emprego de métodos experimentais e quase-experimentais na avaliação de programas públicos. Inicia-se com uma breve apresentação do desenho experimental clássico, seus requisitos e conceitos relacionados, como validades interna, externa e contrafatual. Em seguida, são abordadas as modalidades de desenhos quase-experimentais de avaliação, modelos que flexibilizam os requisitos do experimento clássico. Em duas seções de natureza crítica, apontam-se as limitações éticas, políticas e operacionais dos desenhos experimentais e quase-experimentais na avaliação de programas e, depois, as motivações político-institucionais da resiliência dessa abordagem diante de conhecidos e recorrentes problemas de robustez.

PALAVRAS-CHAVE AVALIAÇÃO • POLÍTICAS PÚBLICAS • EXPERIMENTOS • QUASE-EXPERIMENTOS.

COMO CITAR:

Januzzi, P. de M. (2023). Delineamentos experimentais na avaliação de políticas públicas: Usos e abusos. *Estudos em Avaliação Educacional*, *34*, Artigo e09956. https://doi.org/10.18222/eae.v34.9956

1 Uma versão preliminar deste artigo foi apresentada no XLVI Encontro da Associação Nacional de Pós-Graduação e Pesquisa em Administração (Anpad), em 2022.

DELINEAMIENTOS EXPERIMENTALES EN LA EVALUACIÓN DE POLÍTICAS PÚBLICAS: USOS Y ABUSOS

RESUMEN

El objetivo de este ensayo es discutir el potencial y, sobre todo, los límites del uso de métodos experimentales y casi experimentales en la evaluación de los programas públicos. Comienza con una breve presentación del diseño experimental clásico, sus requisitos y conceptos relacionados, como validez interna, externa y contra fáctica. A continuación, son abordadas las modalidades de los diseños de evaluación casi experimentales, modelos que flexibilizan los requerimientos del experimento clásico. En dos secciones de naturaleza crítica se señalan las limitaciones éticas, políticas y operativas de los diseños experimentales y casi experimentales en la evaluación de los programas, y después, las motivaciones político-institucionales de la resiliencia de este enfoque frente a conocidos y recurrentes problemas de robustez.

PALABRAS CLAVE EVALUACIÓN • POLÍTICAS PÚBLICAS • EXPERIMENTOS • CASI-EXPERIMENTOS.

EXPERIMENTAL DESIGNS IN PUBLIC POLICY EVALUATION: USES AND ABUSES

ABSTRACT

The present essay discusses the strengths and, above all, the limits of using experimental and quasi-experimental methods in evaluating public programs. It begins with a brief presentation of classical experimental design, its requirements and related concepts such as internal, external, and counterfactual validity. Next, it addresses the modalities of quasi-experimental evaluation design, wich relax the requirements of the classical experiment. Two critical sections point out the ethical, political, and operational limitations of experimental and quasi-experimental designs in program evaluation and, subsequently, the political-institutional motivations for the resilience of this approach in spite of well-known and recurrent robustness problems.

KEYWORDS EVALUATION • PUBLIC POLICY • EXPERIMENTS • QUASI-EXPERIMENTS.

Recebido em: 12 DEZEMBRO 2022
Aprovado para publicação em: 11 ABRIL 2023

INTRODUÇÃO

Delineamentos experimentais e quase-experimentais têm um papel destacado na avaliação de projetos, programas e políticas públicas, como bem demonstrado recentemente nas análises de segurança médica e eficácia preventiva das vacinas desenvolvidas – e outros medicamentos usados – para combater os efeitos da pandemia de covid-19. As decisões em programas de saúde pública envolvem riscos muito elevados, que podem ter consequências rápidas, muito além do previsto em situações concretas da realidade social. Felizmente os protocolos de investigação epidemiológica foram desenvolvidos há bastante tempo e desfrutam de forte consenso nas comunidades científica e política no país, garantindo – até há poucos anos – decisões consistentes e responsáveis em políticas de saúde pública no país.

Em outros campos das políticas públicas, como na avaliação educacional e de programas sociais ou mesmo em programas de saúde pública que envolvam intervenções mais complexas – para além de desenvolvimento de vacina, medicamentos ou procedimentos clínicos e terapêuticos –, a pertinência e aplicabilidade desses desenhos avaliativos são ainda objetos de discussão. Para um segmento da comunidade de práticas da avaliação, chamado por Ravallion (2009) de randomistas, os modelos experimentais de pesquisa – e, com alguma concessão, alguns desenhos quase-experimentais – constituiriam o método padrão-ouro para avaliação de programas, o único que atestaria a eficácia e o impacto da intervenção pública (Gertler et al., 2015). Argumenta-se que, garantidos os pressupostos de sua aplicação, esse modelo de pesquisa pode assegurar de forma mais consistente a inferência causal entre a intervenção, suas atividades e produtos e seus efeitos. Pelo que se sugere em manual "oficial" de avaliação de políticas públicas editado pela Casa Civil (Casa Civil da Presidência da República, 2018), evidências produzidas nesse tipo de avaliação seriam mais robustas para formulação e decisão de políticas públicas, cabendo um papel complementar – e menos meritório – aos resultados de avaliações realizadas por meio de outras abordagens. O contexto de primado da austeridade fiscal dos últimos anos e a repercussão pública dos procedimentos de testes de vacina contra a covid-19 certamente têm favorecido essa tendência, legitimando a convicção do modelo experimental como "método mais científico" de avaliação de qualquer iniciativa governamental nas áreas da saúde, educação ou proteção social. Fato é que a crença no método continua vívida e eficiente para capturar corações e mentes nas universidades, no setor público e até mesmo na mídia no país.

Recuperar e sistematizar argumentos acerca dos limites epistemológicos, éticos e operacionais dessa abordagem avaliativa, presentes em várias contribuições de autores nacionais e internacionais, é o objetivo deste ensaio. Partindo do reconhecimento das potencialidades dos métodos experimentais e quase-experimentais na avaliação, procura-se discutir o "lugar" e a "vez", ou contextos e momentos,

do seu uso mais adequado e útil na apreciação de políticas e programas públicos. Busca-se alertar para que o uso, abuso e mau uso dos métodos experimentais na avaliação de programas no país não repliquem a experiência americana dos anos 1960-1970, em que evidências frágeis, mas supostamente "autorizadas", serviram para descontinuar ou deslegitimar programas públicos recém-implantados ou que ainda não tinham tido tempo e recursos suficientes para "entregar" o que se propunham (Patton, 1990, 1997; Rossi, 1987). Se políticas públicas têm relevância e legitimidade política pelos propósitos a que se destinam – isto é, são construções sociopolíticas em dado contexto e sociedade –, é necessário que sejam avaliadas de forma mais consistente e responsável, segundo os modelos sociotécnicos mais apropriados ao processo.

O ensaio inicia-se com uma breve apresentação do desenho experimental clássico de pesquisa e seus requisitos de aleatorização amostral, controle laboratorial/situacional e explicitação de relações causais entre variáveis. Conceitos relacionados como validade interna, externa, contrafatual são também introduzidos. Em seguida, são abordadas as modalidades de desenhos quase-experimentais de avaliação. Esses modelos flexibilizam os requisitos do experimento clássico, como a aleatorização dos grupos tratamento e controle e as condições contextuais de "confinamento". Assim, se, por um lado, nessas modalidades a inferência causal perde a robustez do desenho clássico, por outro, a avaliação do programa se torna exequível, pelo contorno das restrições assinaladas. Finaliza-se o texto com duas seções de natureza crítica, apontando as limitações éticas, políticas e operacionais dos desenhos experimentais e quase-experimentais na avaliação de programas e, em seguida, as motivações político-institucionais da resiliência dessa abordagem frente a tantos e conhecidos problemas de robustez.

UMA BREVE EXPOSIÇÃO SOBRE O USO DO EXPERIMENTO CLÁSSICO NA AVALIAÇÃO DE PROGRAMAS

O modelo experimental clássico na pesquisa epidemiológica tem o objetivo de investigar a estrutura e a intensidade de causalidade entre uma variável-efeito consequente (cura, melhora do estado de saúde) e sua variável-fator determinante (tratamento, medicamento ou vacina). Para isso, é preciso garantir o controle da situação experimental em laboratório e o emprego de dois grupos compostos de forma aleatória a partir de uma mesma população original. Um desses grupos é submetido aos efeitos do novo tratamento que se quer avaliar (grupo tratamento), enquanto o outro não recebe o novo tratamento, mas pode (e deve) ter o tratamento convencional disponível (grupo controle) (Imas & Rist, 2009).

A realização do experimento em laboratórios ou em contexto controlado procura garantir a não interferência de outros fatores que poderiam afetar o estudo,

de modo a deixar mais evidentes os efeitos – ou não efeitos – do novo tratamento. A designação aleatória das pessoas a compor os grupos tratamento e controle busca garantir que, sendo esses indivíduos de uma mesma população-base, as duas amostras são, probabilisticamente, equivalentes ou similares em suas características. São três condições requeridas, portanto, para realização de um experimento: a suposta causa ou fator explicativo (tratamento, medicamento ou vacina) a ser aplicado/testado durante certo período; o controle da situação experimental/laboratorial durante todo o período de análise; e a seleção aleatória das pessoas na composição dos dois grupos – tratamento e controle. Em um estudo em que uma dessas condições é violada, não se pode chamá-lo de experimento.

Um exemplo hipotético, inspirado no teste da vacina Salk contra poliomielite descrito em Cano (2004), pode ajudar a entender a lógica do experimento. Suponha-se, por exemplo, um ensaio clínico – na realidade, um experimento clínico –, em que se pretende avaliar a eficácia de um medicamento – CoroX – contra uma doença nova – CoroV – altamente contagiosa e que, por isso, afetou a saúde de milhares de pessoas e provocou a internação de muitas delas em estado grave nos hospitais. Há, pois, um fator ou variável específico a ser avaliado em termos de seu impacto: se o CoroX cura ou mitiga os efeitos adversos da CoroV. Se há o aval do Comitê de Ética em Pesquisa (CEP) pertinente e a autorização da Agência Nacional de Vigilância Sanitária (Anvisa) para teste do medicamento, pode-se iniciar os preparativos da pesquisa.

Para avaliação do CoroX por meio de um experimento, é necessário garantir o controle situacional/laboratorial: escolher um hospital onde há instalações adequadas para realização da pesquisa e vários pacientes internados em decorrência da CoroV para o teste efetivo do medicamento. Supõe-se que exista um número significativo de pacientes internados, com diferentes níveis de gravidade e tempo de internação em um dado hospital. A última condição para a pesquisa experimental é a criação, por meio de sorteio aleatório, de dois grupos de pacientes: o do tratamento, que receberá regularmente o medicamento CoroX; e o do controle, que será tratado com CoroY, produto com aparência similar mas sem o princípio ativo que se quer testar. Trata-se de um placebo, necessário para que todos os pacientes se sintam em tratamento, já que farta bibliografia médica aponta que fatores psicológicos, como estar sob cuidados médicos, especialmente em uma pesquisa de ponta, favorece a saúde do paciente. Para registro médico e do experimento, são realizados exames em todos os pacientes antes do início do tratamento.

Vale registrar que, com exceção do CoroX, os dois grupos recebem o mesmo tratamento disponível. Pelos princípios da ética médica, não se poderia privar um grupo – controle – do melhor tratamento convencional disponível e já testado. Não se trata, pois, de avaliar o CoroX em relação a nenhum tratamento, mas sim ao

tratamento convencional disponível. A designação aleatória dos pacientes em um dos grupos deve ser, preferencialmente, por meio da estratégia duplo-cego: não só os pacientes não sabem se estão tomando o CoroX ou o placebo CoroY, como também enfermeiros e médicos em contato diário com eles desconhecem a que grupos os pacientes pertencem. Eliminam-se, assim, eventuais conflitos pessoais e vieses de tratamento ou registro da evolução da saúde dos pacientes. Anula-se, inclusive, a eventual tentativa de interferir – por motivos meritórios – no desenho do experimento, como pressionar pela administração do CoroX para todos os pacientes, quando se percebe que o medicamento parece melhorar a saúde de alguns deles. Certamente não são conflitos éticos e humanos fáceis de se lidar, mas equacionados pela ética pragmática dos experimentos em saúde.

Ao longo dos dias, com o avanço do tratamento e acompanhamento da saúde de cada paciente, vai se configurando o momento do exame ou bateria de exames para avaliação da condição de cura, melhora ou não dos pacientes. O momento desse exame "somativo" é crítico: muito cedo, pode não apontar efeitos em processamento; muito tarde, pacientes do experimento podem estar com a saúde em estado ainda mais grave e falecer. Duração e intensidade da dose do medicamento são variáveis críticas a serem analisadas e podem demandar experimentos adicionais.

Em algum momento adequado, é preciso avaliar se o CoroX causou ou não impacto, comparando as medidas-resumo de situação média de saúde dos dois grupos. Conforme ilustrada na Tabela 1, baseada em Rossi et al. (2004), a avaliação da eficácia do CoroX depende da diferença entre T2 e C2, medida a partir do estado de saúde dos dois grupos – tratamento e controle – no momento da avaliação: se T2 é significativamente maior que C2, então o CoroX parece fazer diferença na melhora dos pacientes. No caso mais geral, a avaliação do experimento deve se basear na diferença significativa entre T2-T1 e C2-C1, isto é, se a condição média de saúde do grupo tratamento revelou melhora maior que a do grupo controle (se T2-T1 > C2-C1). Diz-se "significativa" pois se trata de um experimento com amostras da população paciente, não de sua totalidade. Adotado um nível padrão de significância estatística, se a eventual diferença das medidas de melhoria entre grupos é estatisticamente diferente de zero [(T2-T1)-(C2-C1) > 0], haveria forte evidência de que o tratamento com CoroX tem impacto na cura ou melhora da saúde dos pacientes de CoroV. Se, ao contrário, não se observa diferença significativa entre os dois grupos na melhora das condições de saúde, não há evidências de que o medicamento CoroX seja eficaz.

TABELA 1
Avaliação segundo desenho experimental clássico

T1 ----------→ T2	
C1 ----------→ C2	
Medida antes do início do tratamento	Medida após tempo suficiente de tratamento
Onde T: Grupo de tratamento, C: Grupo de controle	
Grupos definidos por designação aleatória e sob controle "laboratorial".	
Se T2 > C2 ou, no caso mais geral, se T2 – T1 > C2 – C1, então há indicações de que o tratamento é eficaz.	

Fonte: Elaboração do autor a partir de Rossi et al. (2004).

A ideia intuitiva subjacente nesse experimento é a de que, como os dois grupos foram expostos às mesmas condições de "laboratório" (internação no mesmo hospital) e são muito parecidos (pois são amostras de um mesmo conjunto inicial de pacientes de CoroV), a eventual diferença entre eles seria consequência do fato de um deles ter tido acesso ao tratamento enquanto o outro não. Eventual diferença na evolução da condição de saúde entre dois grupos similares de indivíduos, mantidos sob mesmo contexto externo, deveria ser imputada, pois existe diferença no tratamento a que os grupos estão submetidos. Essa característica, ou propriedade, do experimento é denominada validade interna.

Mais precisamente, a validade interna de uma pesquisa que busca atribuir relação de causalidade entre duas variáveis é a certeza assegurada por seu desenho metodológico de que a variável independente – ou fator em teste – é responsável pelo impacto observado (positivo ou não) na variável dependente – ou desfecho observado (Imas & Rist, 2009). Como pontuado de forma clara por Cano (2004, p. 29), a "validade interna é o grau de certeza de que o efeito na variável dependente do experimento foi causado pela variável independente do experimento". Trata-se de atestar em que medida foi a causa pesquisada, e não outro fator, que produziu os efeitos observados.

Um desenho de pesquisa goza de alta validade interna quando há grande certeza de que a variável-chave em avaliação – o medicamente CoroX, no exemplo – é o único vetor responsável pelo desfecho observado: melhora ou não dos pacientes com CoroV que foram tratados com o medicamento. Por seu desenho metodológico – aleatorização dos grupos e submissão dos mesmos a idênticas condições externas –, os experimentos gozam de alta validade interna.

Validade interna não é uma propriedade dicotômica (ter ou não ter), mas avaliada em termos gradativos (alta, média e baixa). Assim, outros desenhos não estritamente experimentais de pesquisa, como a comparação de grupos de pacientes com CoroV internados em diferentes hospitais, ou como a comparação de um mesmo grupo de pacientes antes e depois da administração do medicamento, têm menor validade interna. Isso não significa que tais pesquisas – típicas modalidades

denominadas de quase-experimentais – não gozem de nenhuma validade interna, mas sim que a segurança de que o uso do CoroX seja o responsável por eventual melhora observada é menor do que no caso do experimento reportado. Afinal, as equipes dos dois hospitais podem ser diferentes, assim como os equipamentos e instalações, ou ainda que a melhora ao longo do tempo decorra de uma reação positiva dos pacientes ao conjunto de procedimentos terapêuticos e não exclusivamente ao CoroX.

Se validade interna diz respeito à certeza quanto à relação causal intrínseca – ou não – entre duas variáveis, validade externa de uma pesquisa refere-se à segurança com que se pode generalizar seus resultados a outros contextos, populações ou variações metodológicas na medição das variáveis. Apoiando-se novamente em Cano (2004, p. 29), a "validade externa indica a medida em que a inferência causal proposta pelo experimento pode ser generalizada a outros momentos, lugares, populações e formas de medir as variáveis em questão, tanto as independentes quanto as dependentes".

Como já comentado, a pesquisa de avaliação da eficácia do CoroX tem alta validade interna, assegurada pelo desenho experimental adotado. Contudo não se pode assegurar de antemão a validade externa dos seus resultados, de que o CoroX será eficaz no tratamento de qualquer paciente com CoroV, em qualquer hospital, e menos ainda de que tal medicamento se preste à prevenção do contágio com a doença. Afinal, a situação experimental foi bastante particular: um grupo de pacientes internados, em um hospital específico, em um município determinado. Seria eficaz o tratamento com CoroX em pacientes internados em hospitais com equipe médica mais reduzida, menor infraestrutura, sujeitos a problemas nos equipamentos e sobrecarga de atendimento? Em localidades mais frias ou mais quentes, úmidas ou secas, mais adensadas ou não?

Para ampliar a validade externa e assegurar a generalização de resultados de experimentos clínicos, realizam-se estudos multicêntricos, isto é, avalições similares para diversos grupos populacionais, regiões, contextos de tratamento e de gravidade da doença em questão. Procura-se, dessa forma, reproduzir as inúmeras situações concretas e reais encontradas em saúde pública, e não mais no contexto controlado da pesquisa técnico-científica. Idealmente, para assegurar as validades interna e externa do estudo, a amostra de pessoas selecionadas aleatoriamente para os grupos tratamento e controle deveria ser um retrato aproximado dos pacientes internados ou até mesmo indivíduos sãos.

Como sugerido por Gertler et al. (2015), para que seja possível generalizar os impactos identificados em experimentos, a amostra deve ser representativa da população elegível, mas com grupos tratamento e controle formados por meio de processos de aleatorização.

> Uma avaliação de impacto pode gerar estimativas de impacto internamente válidas através da alocação aleatória de tratamento; no entanto, se a avaliação for realizada em uma amostra não aleatória da população, os impactos estimados podem não ser generalizáveis à população de unidades elegíveis. De forma inversa, se a avaliação utilizar uma amostra aleatória da população de unidades elegíveis mas o tratamento não for selecionado de forma aleatória, então a amostra será representativa, mas o grupo de comparação poderá não ser válido. (Gertler et al., 2015, p. 55).

Mas nem sempre é tão simples garantir simultaneamente a validade interna (inferência causal) e a externa (inferência populacional) em experimentos clínicos multicêntricos ou em avaliações experimentais de programas públicos. Assegurar uma pode prejudicar a outra (Cano, 2004). Maior controle experimental para prover maior poder de inferência causal pode significar maior artificialidade do contexto analisado, limitando a generalização dos resultados para contextos mais realísticos; amostras mais amplas e representativas da população podem fragilizar os pressupostos experimentais de controle contextual (e a validade interna da pesquisa).

Na avaliação experimental de programas públicos, o balanceamento entre os dois tipos de validade adquire ainda maior criticidade. Como garantir que um programa bem avaliado em circunstâncias artificialmente construídas possa repetir o êxito em situações normais pelo país afora, com toda heterogeneidade de serviços e equipamentos públicos? Se validade interna é um atributo importante a ser garantido na avaliação, para atestar ou não impacto do programa, "de que serve uma garantia plena sobre uma inferência causal que não pode ser aplicada além do contexto concreto em que foi gerada?" (Cano, 2004, p. 31). É preciso, pois, buscar o equilíbrio entre validade interna e externa na avaliação experimental, ou em qualquer desenho de pesquisa.

AS PRINCIPAIS MODALIDADES DE DESENHOS QUASE-EXPERIMENTAIS DE AVALIAÇÃO

Os desenhos quase-experimentais são efetivamente muito mais comuns do que os delineamentos experimentais na avaliação de programas cujos públicos-alvo sejam pessoas, famílias, empresas ou instituições. Embora com menor validade interna, são mais exequíveis no contexto brasileiro. Na realidade concreta da administração pública, avaliações de impacto de programas públicos com desenhos estritamente experimentais são menos exequíveis do que normalmente se reconhece em alguns manuais de avaliação. Há situações em que se usa impropriamente o termo "experimento natural" na avaliação de um programa quando, na realidade, não se trata

de um experimento clássico como o aqui apresentado, mas sim de uma modalidade quase-experimental. É o caso quando, por questões de cronograma de implementação de um programa, algumas famílias ou indivíduos são atendidos antes de outros, igualmente elegíveis. Supondo-se que o acesso das famílias ou indivíduos ocorra de forma "natural", sem critérios de priorização pessoal, familiar ou regional, o processo emularia uma pseudoaleatorização – e não uma autêntica aleatorização – na formação dos grupos tratamento e controle, o que impede de classificar tal avaliação como estritamente experimental.

Como nos experimentos, a avaliação quase-experimental tem como finalidade verificar se os efeitos observados – impactos – em um grupo de beneficiários ou usuários de um programa decorrem de atividades, produtos e componentes nele desenvolvidos, bem como estimar a dimensão de tal impacto. O que diferencia quase-experimentos dos modelos clássicos é, em geral, a não observação da designação aleatória dos grupos tratamento e controle, ou ainda a falta do controle contextual (e do grupo controle) e, portanto, o relaxamento quanto à interveniência de outros fatores externos na produção de efeitos de um programa (Cano, 2004; Rossi et al., 2004). Uma ou outra situação acaba violando os pressupostos do modelo experimental clássico, afetando a assertividade na atribuição do programa a impactos diferenciais observados nos beneficiários em relação ao grupo de comparação (e não grupo controle, termo usado no caso experimental, como visto).

Nos quase-experimentos os grupos de tratamento e comparação acabam sendo definidos por processos não aleatórios, ou, na melhor das situações, pseudoaleatórios, como na situação anteriormente descrita do "experimento natural". Nesse caso, em função de atrasos ou do calendário de implantação de um programa público, ou ainda pela cobertura deficiente de público-alvo do programa, é possível criar artificiosamente grupos de comparação a partir da análise dos grupos já inseridos. Famílias, pessoas ou empresas elegíveis ao programa, mesmo que não tenham ainda sido inseridas nele, podem servir de unidades amostrais para o grupo de comparação, do qual se pode coletar medidas para fins de comparação com o de tratamento. Cuidados adicionais, com maior ou menor requinte metodológico, podem garantir que os grupos sejam mais próximos, como a aplicação de fatores de calibração PSM (pareamento por escore de propensão), corrigindo o efeito diferencial de características socioeconômicas e demográficas entre os dois grupos, antes e/ou após o período de "tratamento".

De qualquer forma, a não aleatorização na formação dos grupos tratamento e comparação introduz o que a bibliografia da área classifica como viés de seleção nos quase-experimentos. Amostras não aleatórias de uma mesma população têm algum viés, mais flagrante ou pouco transparente. Grupos de tratamento e comparação não aleatorizado têm um viés de seleção. Ainda que os integrantes – famílias,

pessoas e empresas – façam parte de um mesmo público-alvo elegível ao programa, a estratégia de pseudoaleatorização e os recursos paliativos de calibração não conseguem garantir que os dois grupos são equivalentes em todas as possíveis dimensões que poderiam afetar o desfecho quando da participação ou não no programa. Entre pessoas, famílias de um mesmo estrato socioeconômico ou contexto sociocultural ou empresas e instituições de um mesmo ramo ou porte, há aquelas que dispõem de mais recursos de informação, iniciativa, interesse ou emergência para se inscreverem em um programa público. Essas motivações são difíceis de serem objetivamente quantificadas, escapando às possibilidades de calibração e equalização das características dos grupos inscritos e já participantes do programa e aqueles não participantes, mas desejosos de fazê-lo. Fato é que tal motivação pode interferir nos resultados e "inflar" o impacto estimado da participação em um programa, ou, na visão de Gertler et al. (2015), levar à estimação de um falso impacto.

Contudo, fora dos manuais e da modelização acadêmica do mundo, a realidade nunca é tão simplificadamente dicotômica (certo/errado, verdadeiro/falso, etc.). Assim, na realidade complexa das pesquisas de avaliação, da mesma forma como validades interna e externa são medidas em uma escala ordinal (alta, média, baixa), as estimativas de impacto podem ser mais ou menos consistentes ou robustas, sendo esse último termo o mais usado. Porém a robustez da estimativa de impacto não depende apenas das tratativas de evitar o viés de seleção dos participantes no início do desenho avaliativo, mas também da garantia de uma boa realização de todas as etapas da avaliação, algo que não é trivial.

Naturalmente, deve-se fazer o que for possível *ex ante* para minimizar o viés de seleção, respeitando os limites éticos e riscos políticos já mencionados. Da mesma forma, é preciso reconhecer os vieses *ex-post* decorrentes da operacionalização concreta do desenho avaliativo e seus efeitos na composição final dos dois grupos – tratamento e controle. Tampouco pode-se esquecer que experimentos ou quase-experimentos estão sempre sujeitos ao viés de seleção do público-alvo base para a pesquisa avaliativa, isto é, a sua representatividade frente à população potencialmente elegível ao programa. Em termos mais precisos, há o inevitável viés de seleção decorrente da não aleatorização – que afeta a validade interna da avaliação – e o viés de seleção amostral da população elegível – que afeta a validade externa do desenho. Estimativas de impacto efetivo do programa em sua operacionalização cotidiana serão menos ou mais consistentes dependendo da combinação de estratégias de mitigação simultânea dos dois tipos de vieses.

Como já observado, embora possam padecer também de vários dos problemas operacionais já apontados e disporem de menor validade que os experimentos, os quase-experimentos têm maior exequibilidade, já que as restrições éticas e político-institucionais são mais bem equacionadas. Parecem oferecer uma alternativa me-

todológica de avaliação de impacto com equilíbrio entre validade interna, rigor técnico e exequibilidade prática.

Na conceituação proposta por Imas e Rist (2009), há oito modalidades de quase-experimentos, das quais cinco são de especial interesse nessa unidade, classificadas em duas categorias: desenhos de avaliação com múltiplas medidas de um mesmo grupo, sem comparação com outro grupo; e desenhos de avaliação usando grupos similares de comparação (Tabela 2). De modo geral, os quase-experimentos sem grupos de comparação são modelos com validade interna mais baixa do que aqueles que se valem de grupos comparativos para estimar o contrafatual, isto é, para estimar qual seria a situação vivenciada pelo grupo de beneficiários/usuários do programa ao longo do tempo caso não fosse contemplado por ele.

TABELA 2
Experimento e modalidades de quase-experimentos na avaliação de impacto

Modelo experimental clássico	T1 C1 T2 x C2 Impacto estimado = (T2-T1) – (C2-C1)
Quase-experimento sem grupo de comparação	
Desenho com avaliação antes e depois	T1 x T2 Impacto estimado = T2 – T1
Série histórica descontinuada	T1 T2 T3 x T4 T5
Desenho longitudinal	x T1 T2 T3 T4 T5
Quase-experimento com grupos de comparação	
Avaliação pós-tratamento com grupo comparação	T2 x NC2 Impacto estimado = T2 – NC2
Desenho com avaliação pré e pós-tratamento	T1 NC1 T2 x NC2 Impacto estimado = (T2-T1) – (NC2-NC1)

Fonte: Elaboração do autor a partir de Imas e Rist (2009).

Nota: T: grupo tratamento, beneficiário do programa; C: grupo controle (aleatorizado); NC: grupo comparação (não aleatorizado); X: tratamento, programa.

A avaliação com grupos com medidas pré e pós-tratamento é o modelo com desenho mais próximo do modelo clássico experimental. Supondo-se similaridade ou pseudoaleatorização dos grupos tratamento e controle, trata-se de um modelo de avaliação de impacto mais sofisticado e de maior validade interna do que os demais. O impacto pode ser estimado como a diferença das diferenças das medidas pré e pós de cada grupo. O saldo resultante é o impacto [(T2-T1) – (NC2-NC1)]. Tal processo garantiria estimativas de impacto mais consistentes, já que, supostamente, vieses do mesmo grupo seriam eliminados pelas diferenças das medidas de resultados antes e depois.

Embora seja um modelo advogado como desejável – e, para certas comunidades, como o único modelo de avaliação de impacto aceitável quando o experimento não é possível – é menos frequentemente aplicado em larga escala para análise de programas federais no Brasil, pelo custo, complexidade operacional e tempo

de execução entre planejamento, coleta em duas ondas e produção dos resultados. Ainda assim, um dos exemplos a se destacar nesse sentido é a avaliação de impacto do Bolsa Família (Jannuzzi & Pinto, 2013). O delineamento metodológico quase-experimental dessa avaliação foi empregado pelo interesse em captar impactos específicos do programa em várias dimensões socioeconômicas, a partir do levantamento da situação socioeconômica dos grupos populacionais beneficiários e não beneficiários em dois momentos (2005 e 2009). De forma a garantir validades interna e externa à pesquisa e preservá-la de possíveis questionamentos quanto a intervenções programáticas diferenciadas em municípios de sua amostra, a identificação dos endereços – e municípios – dos domicílios entrevistados na primeira rodada da pesquisa não foi repassada pelo Centro de Desenvolvimento e Planejamento Regional (Cedeplar/UFMG), executor da pesquisa de campo, à Secretaria de Avaliação e Gestão de Informação do Ministério de Desenvolvimento Social e Combate à Fome. Tal decisão de preservação do sigilo estatístico, ou melhor, da desidentificação da amostra disponível para o Ministério, foi mantida na segunda edição, conduzida em 2009 pelo consórcio formado pelo International Food Policy Research Institute (IFPRI) com o Instituto Datamétrica.

Tal como na coleta de 2005, a pesquisa investigou um conjunto amplo de dimensões das condições de vida das famílias, passando pelas condições da moradia, caracterização demográfica, educacional, participação no mercado de trabalho, rendimento, percepção sobre os programas sociais e eventos de saúde e antropometria das crianças menores de cinco anos. O desenho da pesquisa permitiu que se trouxessem evidências adicionais sobre impactos do programa, juntando-se a muitas outras produzidas por vários outros estudos. Mas seria equivocado afirmar que, pelo desenho mais sofisticado da pesquisa, esses resultados teriam tido maior repercussão pública ou reconhecimento técnico junto à comunidade de gestores. Não foi isso que ocorreu, pois mesmo outras pesquisas realizadas no programa tiveram baixa apropriação pela mídia e sociedade.

UMA PERSPECTIVA CRÍTICA À MITIFICAÇÃO DOS DESENHOS EXPERIMENTAIS DE AVALIAÇÃO

Ao contrário do que defendem Gertler et al. (2015) e a comunidade epistêmica dos *randomistas* – assim intitulados criticamente por Ravallion (2009) –, os delineamentos experimentais são pouco aplicáveis em casos concretos de políticas públicas por motivos de naturezas ética, política e operacional. Seriam até mesmo desperdício de recursos públicos em muitas situações em que se insiste em aplicá-los, como bem coloca Moral-Arce (2014, p. 40):

> Por desgracia, muchos interessados, a distintos níveles, creen que se puede (y se debe) realizar evaluaciones de impacto rutinaria a todos los programas.

> Paradojicamente, esa insistência em tratar de realizar evaluaciones de impacto de manera sistemática, puede conducir al resultado no deseado de desperdiciar recursos (quem son limitados) por tratar de realizar una evaluacion de este tipo.

Em primeiro lugar, para aplicação do desenho experimental clássico em políticas públicas, há problemas éticos não triviais de como escolher e justificar quem vai ser beneficiário e quem ficará de fora do programa, questão que na prática da pesquisa médica já foi equacionada pelos comitês de ética. Tais instâncias assumem a responsabilidade, risco e legitimidade dessas escolhas em prol do desenvolvimento científico e promoção da cura e saúde, mas asseguram aos pacientes do grupo controle o melhor tratamento então disponível. No campo das políticas públicas, pode se estar diante de uma situação de avaliar o efeito de um programa novo em comparação com nenhum serviço público disponível.

No caso brasileiro, a seleção de beneficiários de programas por meio de sorteio é ainda bastante discutível do ponto de vista ético. Em uma sociedade marcada por grande desigualdade regional, cor/raça, condição socioeconômica e iniquidades sociais como pobreza, fome, trabalho infantil, não parece que a "promessa ética do desenvolvimento científico" consiga calar ou falar mais alto que o "compromisso ético com a dignidade humana". Combate da desigualdade, garantia de direitos sociais básicos e promoção da dignidade humana são valores públicos e princípios presentes em marcos normativos constitucionais e infraconstitucionais no país. Não prover acesso a um programa público de transferência de renda para uma família pobre elegível, a uma cisterna para armazenamento de água para consumo e produção para uma família no semiárido, a uma oportunidade de vaga de curso de qualificação profissional para um trabalhador desempregado, ao atendimento socioassistencial para uma mãe ou família com direitos violados quando se tem possibilidade de atendê-los não parece, de modo algum, compatível com alguma razoabilidade ética-social ou ética-pública. Não se trata, como em experimentos clínicos, de prover acesso a um "tratamento novo" a alguns sorteados (ou sortudos?), assegurando aos demais um "tratamento convencional". Dado o curso intermediário e incompleto em que se encontram as políticas públicas no país, a adoção de um experimento clássico na implementação de um novo programa significa, mais uma vez, reproduzir a desigualdade e iniquidade social da ação pública, provendo a alguns o pacote de benefícios e serviços e a outros, em condição provavelmente similar, nenhum acesso a esses bens ou serviços públicos.

Há certamente argumentos que relativizam tal posicionamento. A uma "ética-pró-dignidade-social" se contraporia uma "ética-pró-eficiência-do-gasto" que, na interpretação dos autores do capítulo "Avaliação de impacto", do manual *Avaliação de políticas públicas*, lançado pela Casa Civil em 2018, conduziria, no médio prazo, à primeira:

> O questionamento sobre se é ético utilizar o método de seleção aleatória também deve ser inserido em contexto amplo sobre o uso eficiente de recursos públicos. Muitas políticas não têm seus impactos avaliados e seguem operando sob a suposição de que alcançam os efeitos desejados. No entanto, há que se reconhecer que só podemos ter confiança nos efeitos de um tratamento quando ele foi colocado sob testes com elevado rigor científico. Muitas vezes, políticas consideradas efetivas não mostram nenhum impacto ou apresentam até impactos contrários ao desejado após serem avaliadas com métodos robustos.
>
> A não avaliação rigorosa dificulta, portanto, averiguar se há desperdícios de recursos públicos que poderiam ser empregados de modo mais eficiente, inclusive direcionados aos mesmos objetivos. Em suma, essa discussão mostra que a crítica de que as avaliações de impacto aleatorizadas são antiéticas deve ser inserida num contexto mais amplo – logo, capaz de considerar as falhas decorrentes de seleção dos beneficiários em cenários alternativos, a realidade de implementação das políticas públicas, suas limitações territoriais e orçamentárias, além do uso eficiente dos recursos públicos avaliado da forma mais robusta possível. (Casa Civil..., 2018, p. 271).

Mas, no mesmo capítulo, um pouco antes da citação acima, os autores reconhecem que há situações em que a seleção aleatória seria pouco justificável e politicamente arriscada.

> Para assegurar que as avaliações aleatorizadas sejam implementadas de modo ético, é necessário que pelo menos dois aspectos sejam atendidos, como a priorização daqueles que possuem necessidades especiais e a transparência do sorteio. Caso existam grupos identificados com necessidades especiais que demandem priorização na política, estes não devem ser excluídos em nome da avaliação de impacto da intervenção. No contexto em que o método de seleção aleatória é empregado, os membros desses grupos devem ser beneficiados sem fazerem parte da aleatorização. Quanto à transparência, é boa prática fazer uso de sorteios públicos após a inscrição, inclusive contando com a presença dos inscritos. Esse tipo de procedimento aberto e claro garante que a seleção seja feita não só oferecendo chances de entrada iguais a todos, mas também evitando que os gestores da política sejam criticados por favorecimento a pessoas ou grupos. (Casa Civil..., 2018, p. 270).

No Brasil, pelo que já sugeria Cano (2004), a concessão de acesso a programas e serviços públicos por meio de "sorteios" parece continuar pouco aceita, seja pela população, seja pelo próprio corpo de servidores públicos. Ele comenta que, em países onde o princípio da avaliação experimental de programas é aceito, isso "custou décadas de persuasão" (p. 24).

Questões éticas à parte, sempre em disputa, em termos operacionais há, de partida, um problema de escala na avaliação experimental: em políticas públicas não se está lidando na dimensão de centenas ou poucos milhares de pessoas em ensaios clínicos ou projetos sociais; são dezenas e centenas de milhares, senão milhões de pessoas a serem potencialmente atendidas. Amostras representativas seriam inevitavelmente grandes e dispendiosas no tratamento e controle contextual. A menos que se consiga, como se registrou há pouco, garantir a validade externa com uma amostra de tamanho justificável em termos de custo, não se deveria empregar o modelo experimental.

Mas o conflito entre validade interna e externa é só mais um dos vários problemas concretos vivenciados na aplicação do modelo experimental para a avaliação de programas públicos. A literatura sobre desenhos experimentais é bastante rica no apontamento das dificuldades e problemas concretos do método que afetam sua validade interna e externa, tais como aqueles indicados por Cano (2004) e Imas e Rist (2009): efeito tendencial anterior; maturação dos efeitos; perda de discriminação da medida de impacto; regressão em direção às médias; mortalidade amostral; interação com outros fatores; interferência com outras políticas; etc. As dificuldades de comunicar os resultados e os elevados custos de realização são outros problemas apontados. Ravallion (2009) acrescenta outro: experimentos avaliativos testam uma hipótese específica e não informam que componente de um programa público está empatando sua plena implementação.

Há também que se apontar um dilema epistemológico no desenho experimental: como assegurar que lógicas complexas de intervenção pública, envolvendo diversos agentes, em contextos tão diferenciados, consigam ser modeladas em variáveis com mensuração regular e precisa no dimensionamento de esforços e efeitos de políticas e programas? Como assegurar que o programa X, seu componente Y ou Z, numa cadeia de intervenção longa e complexa X-Y-W-Z é o fator causal de Z? Ademais, em contextos de implantação simultânea de vários programas públicos há grande probabilidade de que outros programas ou fatores estejam contribuindo na produção do impacto. É o que argumentam Bamberguer et al. (2016) em documento sobre diretrizes de avaliação da Agenda dos Objetivos de Desenvolvimento Sustentável para a ONU Mulheres:

> Em pequenos projetos com um baixo nível de complexidade do programa, arranjos institucionais relativamente simples e um baixo nível de dependência contextual, é possível traçar e avaliar uma relação causal direta entre uma intervenção do programa (por exemplo, água potável, bolsa para meninas frequentarem a escola secundária) e o resultado pretendido (por exemplo, taxas mais baixas de diarreia, taxas mais altas de matrícula de meninas). À medida que o programa se torna mais complexo em termos das três dimensões

> anteriores, o número de variáveis-insumos aumenta (muitas vezes operando de forma diferente em diferentes comunidades ou regiões), o número de possíveis resultados pretendidos e não pretendidos também aumenta e a influência de diferentes partes interessadas e arranjos institucionais torna-se mais complicado, bem como o número de fatores contextuais. Consequentemente, torna-se cada vez mais difícil, ou em muitos casos impossível, determinar relações causais diretas.[2] (Bamberguer et al., 2016, pp. 46-47, tradução nossa).

Fato é que a atribuição de impacto a um programa ou componente dele é cada vez mais difícil em contextos em que políticas públicas são desenhadas para atender a múltiplos objetivos (ainda que com ênfase diferenciada entre eles). Assim, mais do que medir contribuições marginais de programas, não faria mais sentido mensurar os efeitos conjuntos ou combinados?

Há ainda dificuldades operacionais na condução de uma avaliação experimental: como evitar a evasão dos beneficiários, a entrada no grupo tratamento de outros indivíduos que faziam parte do controle e, ainda assim, garantir que os efeitos medidos não sejam afetados? Avaliações experimentais de programas envolvem coleta de dados em campo por pelo menos duas vezes ou ondas, para obter as medidas pré e pós-programa dos grupos de tratamento e controle. Dependendo do espaçamento de tempo entre as duas ondas de levantamento, pode-se perder famílias ou pessoas de um ou outro grupo, pela mudança de endereço, migração para outra localidade, óbito ou recusas em participar (sobretudo entre não beneficiários). Se a perda de unidades amostrais – ou atrito – é diferenciada pelos dois grupos, pode-se enfraquecer a validade interna, já que se poderia levantar hipóteses sobre outras variáveis interferindo no processo. A própria validade externa poderia ser enfraquecida se a amostra final deixa de ter proximidade com o público-alvo de referência do programa. Durante o intercurso das duas ondas, famílias ou pessoas beneficiárias do programa podem deixá-lo ou serem excluídas do mesmo. Ou, como em geral é mais comum, famílias ou pessoas do grupo controle podem se tornar beneficiárias do programa, reduzindo o tamanho amostral do grupo e, novamente, diminuindo a validade interna do experimento. Tal fato interfere no tempo de exposição ao programa, variável-chave na intensidade do impacto. Famílias ou

2 No original: "*In small projects with a low level of programme complexity, relatively simple institutional arrangements and a low level of contextual dependence, it is possible to trace and evaluate a direct causal relationship between a programme intervention (e.g., drinking water, scholarships for girls to attend secondary school) and the intended outcome (e.g., lower rates of diarrhea, higher rates of girl's enrolment). As programmes become more complex in terms of the three previous dimensions, the number of inputs increases (often operating differently in different communities or regions), the number of intended and unintended outcomes also increases, and the influence of different stakeholders and institutional arrangements becomes more complicated, as well as the number of contextual factors. Consequently, it becomes increasingly difficult, or in many cases impossible, to determine direct causal relationships*".

ISSN: 0103-6831 • e-ISSN: 1984-932X

indivíduos beneficiários de um programa podem ter acesso facilitado a outros programas, por decisão de formuladores ou pelo maior conhecimento que passam a ter sobre políticas públicas. Mais um revés para a validade interna. Na vida real, o controle contextual na avaliação de programas é muito menos efetivo do que o possível no teste laboratorial de medicamentos e vacinas. Enfim, pode ser que a amostra da segunda onda de coleta de dados apresente grupos tratamento e controle com vieses significativos, eliminando todas as garantias de validades interna e externa que a seleção aleatória proporcionava no início da avaliação.

Encontram-se também desafios metodológicos na escolha de qual é a melhor medida para captar a dimensão impactada. O que se supõe como dimensão impactada guarda, pelo desenho lógico do programa, vinculação estreita com as ações, produtos e serviços dele? O efeito a ser medido é um resultado concreto do programa, assegurável tão somente pelo programa ou um efeito desejável ou potencial, que requer outros pressupostos ou ações não previstos no desenho do programa? O impacto deve ser medido sobre os beneficiários, nas suas famílias, comunidade a que pertencem ou, ainda, na sociedade?

Mesmo que todos esses problemas fossem contornáveis, restaria mais um de natureza prática: se os efeitos potenciais do programa, tal como medidos em uma determinada variável, não forem elevados, as amostras de beneficiários atendidos e do grupo controle teriam de ser consideravelmente grandes para que os testes estatísticos possam ser aceitos sem hesitação (Rossi et al., 2004). Os indicadores antropométricos computados na avaliação de impacto do Bolsa Família são uma ilustração clara nesse sentido: não sendo a transferência monetária mensal do programa um valor tão significativo a ponto de trazer mudanças expressivas na dieta familiar, as diferenças de altura e peso entre crianças de famílias do programa e aquelas não pertencentes a ele são pouco significativas. Fosse a amostra maior, provavelmente o impacto do programa teria sido constatado não apenas no índice de massa corporal, mas também na altura média das crianças, indicador mais refinado de análise (Jaime et al., 2014).

A respeito do uso de testes de significância na pesquisa social, Sawyer e Peter (1983) comentam que é ilusório considerá-los como procedimentos estritamente objetivos quando, ao pesquisador, é dada a oportunidade de alterar, a seu gosto, os diversos parâmetros que podem determinar a significância ou não de uma associação entre variáveis. Aumentar ou diminuir o tamanho da amostra à qual se aplica o teste, decidir por um teste mono ou bicaudal, alterar *post hoc* o nível de significância são procedimentos subjetivos – e não necessariamente condenáveis – que estão por detrás da aparente formalidade e precisão matemática dos testes de significância. Há mesmo um movimento significativo de pesquisadores que propõe abolir os

testes de significância, por entender que sugerem uma "robustez" nos achados que efetivamente não teriam como prover (McShane et al., 2019).

Há na bibliografia aplicada muitas soluções *ad hoc* para vários desses problemas concretos em campo, desde a calibração de amostras dos dois grupos, passando pelas variáveis instrumentais, até outras soluções econométricas não tão consensuais. Mas é preciso estar atento para os limites existentes na adoção de procedimentos para correção de não respostas, desbalanceamento de amostras, perdas amostrais nos levantamentos de campo. O alegado rigor diferencial do modelo experimental frente a outros desenhos avaliativos de programas pode ter ficado na proposta da avaliação, perdendo-se gradativamente a cada passo concreto na execução. Talvez, em alguns casos, a solução metodológica mais adequada – e mais honesta do ponto de vista científico – seja reconhecer que a avaliação deixou de ser experimental e tornou-se um quase-experimento.

COM TANTOS PROBLEMAS DE ROBUSTEZ, POR QUE ESSES DESENHOS AVALIATIVOS RESISTEM AO TEMPO E CONTEXTOS?

A crença de que avaliações de impacto experimental ou suas variações constituem-se no padrão-ouro é reforçada, em um círculo "autorreferenciado", pelos bancos multilaterais de fomento e outras comunidades de financiadores de projetos sociais. Tais instituições, em geral constituídas por equipes com formação acadêmica marcadamente disciplinar e positivista, com pouco conhecimento de desenho e prática de gestão de programas, reforçam essa lógica perversa professada por essa comunidade epistêmica: só colocam recursos em iniciativas em que o gestor se compromete a seguir a cartilha prévia da avaliação de impacto, qualquer que seja a natureza da intervenção, viabilidade operacional do delineamento ou os princípios éticos a obedecer.

É o que La Rovere (2014) discute no contexto de avaliação de políticas ambientais, em que a investigação de contribuições marginais de iniciativa na área e a separação de unidades investigadas em amostra de tratamento e de controle são operacionalmente inviáveis. E desnuda como funciona o círculo de financiamento-método-financiamento de projetos e programas:

> . . . a pressão proveniente de fontes múltiplas (doadores e fóruns de avaliação) em direção ao maior rigor técnico percebido, alcançável por meio de abordagens quantitativas e atribuição de efeitos, está sendo reaplicada na avaliação de impacto e nos profissionais de avaliação. Esta demanda é estimulada (ou muitas vezes reforçada) por grandes doadores que insistem que uma abordagem quantitativa é a única crível.
>
> Esses doadores influentes estão quase sempre localizados nos mesmos lugares (ou seja, países, cidades e, muitas vezes, círculos intelectuais) que as

> instituições acadêmicas onde essas ferramentas estão sendo promovidas.[3] (La Rovere, 2014, p. 285, tradução nossa).

Nessa mesma linha, uma pesquisa extensa que se dedicou a entender por que os modelos experimentais tiveram uma segunda onda de ascensão a partir dos anos 2000, depois do descenso na década de 1970, explica que:

> Nosso argumento é que a expansão contemporânea dos experimentos aleatórios clássicos só pode ser entendida no contexto de duas transformações independentes que então se ligaram uma à outra por afinidades eletivas. Por um lado, o campo da ajuda externa foi profundamente transformado com a entrada de um novo conjunto de atores: grandes fundações privadas com uma ambição global e um novo estilo gerencial (o denominado "filantropo-capitalismo". . . . Por outro lado, o campo da a economia do desenvolvimento foi transformada pelo surgimento da economia comportamental, com sua ênfase em modelos cognitivamente plausíveis de atores humanos e o uso de cutucadas para direcionar atores em uma direção economicamente racional. . . . O sucesso dos randomistas deve ser entendido em função de sua capacidade de explorar essa afinidade entre as duas transformações, com experimentos servindo como "elo" entre os "interesses conectados" da profissão de economia e o campo da ajuda ao desenvolvimento. . . . O sucesso de randomistas, veremos, dependiam de sua capacidade de efetivamente mudar o que era um "experimento de campo " e o que significa avaliar a política de desenvolvimento, uma mudança para a qual a privatização da ajuda externa proporcionou um contexto favorável, enquanto a economia comportamental proporcionou um modelo conceitual e um conjunto de ferramentas.[4] (Leão & Eyal, 2016, p. 3, tradução nossa).

3 No original: "*Yet pressure arising from multiple sources (donors and evaluation fora) towards the perceived higher rigour achievable through quantitative approaches and attribution is being reapplied on impact assessment and evaluation practitioners. This demand is stimulated (or often enforced) by major donors insisting that a quantitative approach is the only credible one. These influential donors are almost always located in the same places (i.e. countries, cities and often intellectual circles) as the academic institutions where such tools are being promoted*".

4 No original: "*Our argument is that the contemporary expansion of RCTs can only be understood in the context of two independent transformations that then became linked to one another in elective affinity. On the one hand, the field of foreign aid has been profoundly transformed by the entry of a new set of actors: large, private foundations with a global ambition and a new managerial style (the so-called 'philantrocapitalism' . . . . On the other hand, the field of development economics has been transformed by the rise of behavioral economics, with their emphasis on cognitively plausible models of human actors and the use of 'nudges' to channel actors in an economically rational direction . . . . The success of randomistas must be understood as a function of their capacity to exploit the elective affinity between these two transformations, with RCTs serving as the 'hinge' between the 'inked ecologies' of the economics profession and the field of development aid . . . . The success of randomistas, we shall see, depended on their capacity to effectively change what is a 'field experiment' and what it means to evaluate development policy, a change for which the privatization of foreign aid provided a hospitable ecology, while behavioral economics provided a conceptual model and a set of tools*".

Os autores atribuem, pois, menos às propaladas virtudes de robustez do método experimental e muito mais à conjunção de interesses – afinidades eletivas[5] – de uma nova escola de pensamento econômico em um novo contexto de financiamento de projetos de desenvolvimento. Os filantropo-financiadores de projetos sociais, por formação ou necessidade de procedimentos "objetivos", estariam ou teriam se convencido que a avaliação experimental atenderia bem a esses objetivos. A comunidade de economistas dessa nova escola comportamental viu a oportunidade de resgatar o ferramental produzido há algumas décadas e reembalá-lo como a mais nova e robusta metodologia de avaliação, o padrão-ouro a ser empregado, em detrimento das demais abordagens. Ocorre que, segundo os autores, essa abordagem avaliativa havia deixado de ser usada exatamente por sua inadequação aos problemas e complexidade de implementação dos programas públicos de então.

A mitificação desse desenho na avaliação de programas deve-se, em alguma medida, à origem dos estudos avaliativos na investigação de programas nas áreas de educação e saúde pública, em que tais modelos podem ser viabilizados mais concretamente – pelas condições de simulação de "laboratório" em salas de aula ou pela tradição dos ensaios de tratamento clínico de doenças (Leeuw, 2010). Outro fator explicativo é a hegemonia circunstancial dos modelos quantitativos das ciências naturais, no debate sobre a cientificidade dos métodos de pesquisa na pesquisa social americana na década de 1960 – momento de expansão dos estudos avaliativos naquele país (Jannuzzi, 2018). De fato, o livro clássico *Experimental and quasi-experimental designs for research*, publicado em 1966, teve forte influência na formação e práticas de avaliação nas universidades americanas, em parte pela elegância e validade interna para análises de causalidade. Nesse período, apesar das advertências sobre as dificuldades de replicação das condições de controle laboratorial no contexto de operação dos programas sociais, "a elegância e a precisão do método experimental levaram a maioria dos avaliadores de programa a vê-lo como ideal" (Worthern et al., 2004, p. 116).

As críticas que se seguiram nas décadas posteriores sobre aspectos éticos, factibilidade operacional e poder de generalização dos resultados de desenhos experimentais – e suas aproximações quase-experimentais seja na pesquisa acadêmica, seja na pesquisa de avaliação de programas –, a incorporação de avaliadores provenientes das várias disciplinas das ciências sociais – antropólogos, sociólogos, comunicólogos, etc. – e a formalização mais rigorosa de abordagens de investigação mais qualitativas, mais adequadas aos problemas complexos e pouco estruturados da realidade social, acabaram por consolidar a percepção, na comunidade de

5 De forma simplificada, "afinidade eletiva" é um conceito sociológico empregado com recorrência para ilustrar situações de convergência de interesses entre atores ou correspondência estrutural entre os mesmos.

avaliadores nos EUA, de que os estudos avaliativos requerem certo ecletismo metodológico, integrando métodos quantitativos e qualitativos.

Desenhos experimentais e quase-experimentais de avaliação de políticas públicas envolvem não só questões éticas de difícil contorno – como a escolha de quem faz parte dos grupos controle e tratamento – mas também problemas de operacionalização nada triviais – como o "isolamento" dos dois grupos ao longo do tempo e a garantia de "isonomia" das demais condições contextuais. Em uma remissão a visões já ultrapassadas acerca da produção do conhecimento científico, se autodeclaram como "politicamente neutros" e "cientificamente atestados". Esquecem-se que a atribuição (ou deslegitimação) dos efeitos identificados em uma população aos componentes de um programa depende de muitas escolhas quanto aos testes estatísticos, níveis de significância, características e tamanho de amostras, dos pressupostos com relação às propriedades de distribuição dos dados. Não se encontra em muitos desses trabalhos a discussão sobre poder estatístico dos testes usados ou sobre a análise de resíduos após a estimação de parâmetros de modelos. Ainda menos comuns são análises mais exaustivas sobre os potenciais vieses introduzidos na estimação do sentido e intensidade do impacto (ou não impacto) pelas calibrações dos grupos tratamento e controle pela técnica *propensity score matching*. A impressão que se tem em muitos trabalhos é que são aceitas ou rejeitadas hipóteses de relação causal mais por convicções do que efetivamente por "provas cabais".

Ademais, a escolha da amostra de análise, que permite driblar os imperativos éticos ou operacionais na separação entre grupos de tratamento e de controle, é por vezes muito particular, padecendo da crítica que imputam às amostras de outros estudos que reputam com desenhos "menos científicos". Isto é, também são amostras com viés de representatividade (e por que não de seleção?), em que a validade externa pode estar comprometida, a fim de garantir as condições de validade interna (aleatoriedade ou quase-aleatoriedade na definição dos dois grupos). Se as amostras selecionadas potencializam a validade interna do desenho da pesquisa avaliativa (e a relação de atribuição entre causa e efeito), há que se reconhecer que muitas vezes isso se faz em detrimento da representatividade do público-alvo dos programas e da realidade dura e concreta da implementação de programas em ambientes complexos.

Desenhos experimentais têm certamente relevância e aplicação na avaliação de políticas públicas. Seu emprego como estratégia de avaliação de impacto requer uma série de reflexões acerca do momento, custos e expectativas de resultados. Classificá-los como padrão-ouro é não só equivocado em termos de prescrição geral frente às diferentes perguntas avaliativas possíveis para uma dada política pública, como também pouco responsável pelas implicações éticas e políticas requeridas para sua realização.

Não existe um só modo de fazer ciência ou um único método "padrão-ouro" de produção e legitimação de conhecimento. Também não existem "verdades absolutas e incontestes" nas ciências, quanto mais nas ciências sociais e na avaliação (Gussi & Oliveira, 2017). Tais "verdades" são o que se procura em algumas religiões; nas ciências sociais e avaliação buscam-se achados e interpretações consistentes sobre o que se analisa, ou ainda, em uma perspectiva "latouriana", fazer ciência é construir narrativas persuasivas e convincentes sobre os achados analisados, reconhecidas como legítimas e razoáveis pelas comunidades epistêmicas a que se pertence.

Posição crítica semelhante tem Chianca (2015, p. 28), exposta em interessante trabalho, do qual vale citar na íntegra suas considerações finais:

> a. Dizer que um único método é o padrão ouro é como dizer que um único medicamento é o melhor que existe. Você precisa se perguntar sempre se ele é o melhor para que tipo de problema de saúde.
>
> b. O verdadeiro padrão ouro é aquele que consegue estabelecer uma argumentação causal que seja consistente e dentro da necessidade de precisão requerida pelo contexto avaliativo, com base em evidências corretas e robustas que, ao mesmo tempo, apoiem e testem criticamente essa argumentação.
>
> c. Não se quer escolher um único método. De longe, os melhores desenhos avaliativos ou de pesquisa usam o princípio do multiplismo crítico . . . que significa o emprego de uma combinação de métodos, sendo que os pontos fortes de uns compensam os pontos fracos de outros e vice-versa. Todo método tem suas limitações. Portanto, basear-se em apenas um deles é uma prática inadequada.
>
> d. Há um último ponto crítico que precisa ser considerado. Não se deve escolher apenas o método ou o conjunto de métodos que parece tecnicamente mais adequado para uma situação. Trabalha-se em situações de vida real, com limitações de recursos. Portanto, é preciso escolher métodos de inferência causal mais custos-efetivos para responder de maneira suficientemente robusta às nossas perguntas avaliativas.

Conforme defendido por Weiss (1998), Vaitsman e Paes-Sousa (2009), Batista e Domingos (2017) e diversos outros autores aqui citados, como Imas e Rist (2009), Moral-Arce (2014) e Bamberguer et al. (2016), as avaliações de políticas e programas têm muito a ganhar em qualidade e consistência com abordagens complementares de métodos quantitativos, qualitativos, experimentais e quase-experimentais.

CONSIDERAÇÕES FINAIS

Experimentos são pouco realizáveis na realidade brasileira, mesmo com a pressão de duas décadas de organismos multilaterais de fomento para torná-los mais regulares na avaliação de programas. Questionamentos éticos e políticos acerca de um eventual acesso a programas públicos por meio de sorteio ainda parecem instransponíveis para a sociedade e gestão pública no país. No quadro de desigualdades e iniquidades estruturais no Brasil, não parece sensato render-se às veleidades metodológicas quando há parâmetros substantivos de elegibilidade e critérios republicanos de priorização social para garantir não apenas o acesso a programas, mas também a direitos assegurados na Constituição.

Quase-experimentos têm potencial para contornar parte desses problemas éticos e políticos, mas também podem padecer de dificuldades de realização operacional, que comprometem as validades interna e externa inicialmente planejadas. Os dois casos, em geral, implicam custos elevados de campo, estrutura de supervisão bem qualificada e bom tempo de realização (entre as ondas de levantamento). Estudos longitudinais com integração de registros de programas e cadastros públicos podem ser alternativas menos onerosas em tempo e operação de campo, mas dependem da qualidade dos dados disponíveis, variáveis-chave para vinculação física dos registros de beneficiários e acesso efetivo aos dados, custodiados por diferentes agentes públicos, nem sempre – e justificadamente – dispostos a compartilhar informação pessoal ou familiar de cidadãos sem uma clara e meritória razão para tanto.

A modelagem de bases integradas dos dados provenientes de registros administrativos e cadastros públicos, criados para gestão de políticas e programas setoriais, é uma alternativa a essas duas modalidades (Jannuzzi, 2016). Asseguradas a qualidade, atualidade e especificidade da informação registrada nessas fontes de dados, tal estratégia permitiria construir modelos com muitas possibilidades de avaliação comparativa – ou pseudoaleatorização – de situações factuais e contrafatuais, "tratamentos" e "controles", interação maior ou menor de programas e de contextos diferenciados dos públicos atendidos ou de agentes operadores dos programas. Delineamentos quase-experimentais *ex post* também seriam passíveis de serem simulados por meio dessa estratégia metodológica. Ademais, e talvez sua principal vantagem comparativa, tal estratégia permite, adicionalmente, o pareamento longitudinal das unidades de análise em painéis com extensão histórica ou periodicidade bem mais flexíveis e interessantes para as análises de efeitos do "tempo ou regularidade de exposição ao programa social".

Como todo método de pesquisa empregado na avaliação de programas, há potencialidades e limitações que precisam ser analisadas caso a caso. Mas é fato que, nesses delineamentos avaliativos, programas públicos não podem continuar

sendo modelizados com uma parametrização *dummy* – 0 ou 1 em uma variável sinalizadora – sem considerar as características básicas de seu desenho, em especial a intensidade e tempo dos beneficiários aos serviços e benefícios dos programas, isto é, a magnitude da "dose e exposição ao tratamento".

As considerações aqui sistematizadas acerca de métodos experimentais e correlatos deveriam, de alguma forma, levar a uma reflexão crítica sobre a posição assumida por parte da comunidade epistêmica e de práticas na avaliação acerca da propalada superioridade ou robustez desses modelos em relação a outros na avaliação de programas. A robustez de uma avalição não é assegurada pela sofisticação técnica, elegância formal ou qualidades teóricas intrínsecas do método a ser empregado. A robustez de uma avaliação depende da adequação do método ao problema em questão, da consistência técnica com que o método é efetivamente empregado no início, meio e fim do processo avaliativo, da amostra, coleta de dados e supostos utilizados na análise. A robustez de uma avaliação depende da consensualidade e transparência das escolhas técnicas diante dos desafios metodológicos que inexoravelmente surgem em problemas complexos, quando se quer resolvê-los segundo perspectivas compreensivas de tratamento e análise e não conforme conveniências e veleidades do método de preferência. A robustez de uma avaliação depende da honestidade intelectual com que supostos e pressupostos acerca do programa, seu mérito e resultados obtidos por uma perspectiva metodológica são colocados à prova mediante a triangulação de outros achados, obtidos por outros métodos, sujeitos e perspectivas interpretativas.

Enfim, a robustez de uma avaliação não está associada a um suposto método padrão-ouro, mas sim a uma postura técnico-científica padrão diamante, esclarecida, plural e vigilante, que reconheça a natureza contingencial e limitada do conhecimento sobre a realidade complexa das demandas públicas, problemas coletivos e ações governamentais desenhadas para atendê-las ou mitigá-las.

REFERÊNCIAS

Bamberguer, M., Segone, M., & Tateossian, F. (2016). *Evaluating the Sustainable Development Goals with a "no one left behind" lens through equity-focused and gender-responsive evaluations*. UN Women.

Batista, M., & Domingos, A. (2017). Mais que boas intenções: Técnicas quantitativas e qualitativas na avaliação de impacto de políticas públicas. *Revista Brasileira de Ciências Sociais, 32*(94), Artigo e329414. https://doi.org/10.17666/329414/2017

Cano, I. (2004). *Introdução à avaliação de programas sociais*. FGV.

Casa Civil da Presidência da República. (2018). *Avaliação de políticas públicas: Guia prático de análise ex post* (vol. 2). Casa Civil da Presidência da República.

Chianca, T. (2015). Um modelo alternativo ao estudo experimental para inferir causalidade em avaliações do impacto de projetos sociais. *Revista Brasileira de Monitoramento e Avaliação, 9,* 16-29. http://dx.doi.org/10.4322/rbma201509003

Gertler, P., Martínez, S., Premand, P., Rawlings, L. B., & Vermeesch, C. M. J. (2015). *Avaliação de impacto na prática*. Banco Mundial.

Gussi, A. F., & Oliveira, B. R. (2017). Discutindo paradigmas contra-hegemônicos de avaliação de políticas públicas. *Anais do 1. Encontro Nacional de Ensino e Pesquisa do Campo de Públicas*. UFC.

Imas, L. G. M., & Rist, R. (2009). *The road to results: Designing and conducting effective development evaluations*. World Bank.

Jaime, P. C., Vaz, A. C. N., Nilson, E. A. F., Fonseca, J. C. G., Guadagnin, S. C., Silva, S. A., Sousa, M. F., & Santos, L. M. P. (2014). Desnutrição em crianças de até 5 anos beneficiárias do programa Bolsa Família: Análise transversal e painel longitudinal de 2008 a 2012. *Cadernos de Estudos Desenvolvimento Social em Debate,* (17), 49-61.

Jannuzzi, P. de M. (2016). *Monitoramento e avaliação de programas sociais: Uma introdução aos conceitos e técnicas*. Alínea.

Jannuzzi, P. de M. (2018). Mitos do desenho quase-experimental na avaliação de programas. *Nau Social, 9*(16), 76-90. https://doi.org/10.9771/ns.v9i16.31419

Jannuzzi, P. de M., & Pinto, A. (2013). Bolsa Família e seus impactos nas condições de vida da população brasileira: Uma síntese dos principais achados da pesquisa de avaliação de impacto do Bolsa Família II. In M. Neri, & T. Campello (Orgs.), *Programa Bolsa Família: Uma década de inclusão e cidadania* (pp. 179-192). Ipea.

La Rovere, R. (2014). The consultative group on international agricultural research approach to impact evaluation on environment and natural resources management. In J. I. Uitto (Org.), *Evaluating environment in international development* (pp. 277-288). Routledge.

Leão, L. S., & Eyal, G. (August, 2016). Experiments in the wild: A historical perspective on the rise of randomized controlled trials in international development. *Comparative Historical Sociology Mini-Conference*. American Sociological Association, Seattle, USA.

Leeuw, F. L. (2010). On the contemporary history of experimental evaluations and its relevance for policy making. In O. Rieper, F. L. Leeuw, & T. Ling (Eds.), *The evidence book: concepts, generation, and use of evidence* (Comparative polyce evaluation, vol. 15, pp. 11-26). Routledge.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. F. (2019). Abandon statistical significance. *The American Statistician, 73,* 235-245. https://doi.org/10.1080/00031305.2018.1527253

Moral-Arce, I. (2014). *Elección del método de evaluación cuantitativa de una política pública: Buenas prácticas en América Latina y la Unión Europea*. EuroSocial.

Patton, M. Q. (1990). *Qualitative evaluation and research methods.* Sage.

Patton, M. Q. (1997). *Utilization-focused evaluation: The next century text*. Sage.

Ravallion, M. (2009). Should the randomists rule? *The Economists' Voice, 6*(2). https://doi.org/10.2202/1553-3832.1368

Rossi, P. (1987). The iron law of evaluation and other metallic rules. In J. Miller, & M. Lewis, *Research in social problems and public policy* (vol. 4, pp. 3-20). Jai Press.

Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2004). *Evaluation: A systematic approach*. Sage.

Sawyer, A. G., & Peter, J. P. (1983). The significance of statistical significance tests in Marketing Research. *Journal of Marketing Research, 20*(2), 122-133. https://doi.org/10.2307/3151679

Vaitsman, J., & Paes-Sousa, R. (2009). Avaliação de programas e transparência da gestão pública. In C. Franzese, C. Anjos, D. Ferraz, F. L. Abrucio, G. N. Cheli, G. M. B. P. Melo, J. Vaistman, J. L. D. Nehmé, L. Santoni, M. G. da Silva, M. Rubio, P. Yanes, S. Nahas, P. de M. Jannuzzi, & R. Paes-Sousa, *Reflexões para Ibero-América: Avaliação de programas sociais* (pp. 11-23). Enap.

Weiss, C. I. (1998). *Evaluation research*. Prentice Hall.

Worthern, B. R., Sanders, J. R., & Fitzpatrick, J. L. (2004). *Avaliação de programas: Concepções e práticas*. Gente.