REFORMAS ENCARADAS COMO EXPERIMENTOS*

*Donald T. Campbell***

Os Estados Unidos e outras nações modernas deveriam estar preparadas a dar uma abordagem experimental à reforma social, abordagem na qual novos programas objetivando solução de certos problemas sociais específicos seriam tentados. Esse procedimento possibilitaria verificar a eficácia ou não dos programas examinados a fim de que sejam mantidos, imitados, modificados ou descartados, com base em sua eficácia manifesta, analisada através de múltiplos critérios disponíveis. Nosso preparo para isso é indicado pela inclusão de cláusulas específicas de avaliação de programas na primeira onda de legislação da "Great Society" e pelas propostas em curso no Congresso americano para o estabelecimento de "indicadores sociais" e "bancos de dados" socialmente relevantes. Pelo fato de já há algum tempo termos tido boas intenções nesse sentido, muitos podem achar que já atingimos esse estágio, que já estamos prosseguindo ou suspendendo programas com base numa eficácia avaliada. Um dos temas deste artigo é mostrar que isso não ocorre, que muitos programas de melhorias terminam sem *nenhuma* avaliação interpretável (Etzioni, 1968; Hyman & Wright, 1967; Schwartz, 1961). Precisamos examinar diligentemente as origens dessa condição e esquematizar maneiras de vencer as dificuldades. Este trabalho é um esforço preliminar nesse sentido.

Muitas das dificuldades estão nas intransigências do ambiente de pesquisa e na presença freqüente de sedutoras ciladas de interpretação. A maior parte deste artigo será dedicada a esses problemas. Acontece, no entanto, que as poucas soluções disponíveis dependem de decisões administrativas corretas no iniciamento e na execução do programa. Tais decisões são tomadas na esfera política e envolvem riscos políticos que muitas vezes são suficientes para explicar a falta de uma avaliação criteriosa dos efeitos. A remoção dos administradores de reformas para fora do foco político parece ser tanto improvável como indesejável, ainda que fosse possível. O que é essencial, em vez disso, é que o orientador de pesquisa do cientista social compreenda as realidades políticas da situação e que sua ajuda seja no sentido de estimular a criação de uma demanda pública de avaliações sérias, contribuindo assim para as invenções políticas que reduzem o perigo de avaliações honestas e educando futuros administradores nos problemas e possibilidades.

Por este motivo, há também uma tentativa neste artigo de considerar o meio político da avaliação de programas e de oferecer sugestões de posturas políticas que possam favorecer uma abordagem verdadeiramente experimental à reforma social. Embora tais considerações sejam distribuídas no curso deste trabalho como um tema de menor interesse, parece conveniente começar com algumas idéias gerais de natureza política.

1. VULNERABILIDADE POLÍTICA POR SE CONHECEREM RESULTADOS

Um dos aspectos mais característicos da situação atual relaciona-se ao fato de que *reformas específicas são preconizadas como se o seu sucesso fosse certo.* Por essa razão, o conhecimento dos resultados tem implicações políticas imediatas. Dada a dificuldade inerente de se

*Publicado originalmente em inglês, sob o título Reforms as experiments. *American Psychologist*, v. 24, n. 4, p 409-29, Apr., 1969. A preparação do original deste trabalho foi subsidiada pela National Science Foundation. Várias versões foram apresentadas nas seguintes formas: como uma conferência para o Fundo de Ex-Alunos da Universidade de Northwestern, em 24 de janeiro de 1968; para a Seção de Psicologia Social da Sociedade Britânica de Psicologia, em Oxford, em 20 de setembro de 1968; para o Congresso Internacional de Psicologia Social, em Praga, em 7 de outubro de 1968 (sob título diferente); e para vários outros grupos. Traduzido para o português pelo Prof. Fábio Luiz Mariotto, da Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas. O tradutor agradece a colaboração de Mario Mariotto.

**Professor do Departamento de Psicologia da Northwestern University, em Evanston, Illinois, EUA.

conseguir melhorias significativas com os meios usualmente fornecidos e dada a discrepância entre promessas e possibilidades, a maior parte dos administradores prefere, sensatamente, restringir as avaliações àqueles resultados que conseguem controlar, especialmente no que se refere a resultados a serem anunciados ou divulgados pela imprensa. A ambigüidade, a falta de bases verdadeiras de comparação e de evidência concreta conspiram para aumentar o controle do administrador sobre o que é dito, ou, pelo menos, para reduzir o impacto da crítica no caso de fracasso real. Há segurança sob o manto da ignorância. Além dessa conjunção de promessa e gestão, há outra fonte de vulnerabilidade na circunstância de que os fatos relevantes para a avaliação de um programa podem também ser utilizados para se questionar a eficiência geral e até a honestidade dos administradores. A acessibilidade de tais fatos ao público reduz a intimidade e a segurança de alguns administradores.

Mesmo quando existe um compromisso ideológico para uma avaliação séria da eficiência organizacional ou para uma organização científica da sociedade, esses dois perigos levam à impossibilidade de avaliação realística de experimentos organizacionais. Se o sistema político e administrativo comprometeu-se antecipadamente à correção e eficácia de suas reformas, ele não pode tolerar o reconhecimento do fracasso. Para sermos realmente científicos é preciso que sejamos capazes de experimentar e que possamos preconizar sem aquele excesso de compromisso que nos torna cegos ao teste da realidade.

Esse transe, favorecido pela apatia pública e pela corrupção deliberada, pode vir, a longo prazo, a impedir uma abordagem verdadeiramente experimental para a melhoria social. Mas nossas necessidades e esperanças de uma sociedade melhor exigem que façamos o esforço. Há alguns sinais de esperança. Nos Estados Unidos conseguimos obter índices de custo de vida e desemprego que, embora imperfeitos, têm embaraçado os governos que os publicam. Temos conseguido efetuar recenseamentos que reduzem o número de deputados que um estado tem no Congresso. Esses são motivos de otimismo, embora a morosidade corrupta dos governos estaduais em seguir suas próprias constituições para a revisão de distritos legislativos ilustra o problema.

Uma mudança simples de postura política que reduziria o problema seria a de passar de uma preconização de uma reforma específica para a preconização da seriedade do problema e daí para a preconização de uma persistência em esforços alternativos de reforma, caso o primeiro falhasse. A posição política seria: "Este problema é sério. Propomo-nos a adotar a 'Política A' numa base experimental. Se após cinco anos não houver ocorrido uma melhora significativa, mudaremos para a 'Política B'." Por tornar explícito que a solução dada ao problema foi somente uma dentre as que o administrador ou partido poderia preconizar em sã consciência e por ter já pronta uma alternativa plausível, o administrador teria condições para uma avaliação honesta de resultados. Resultados negativos como o do fracasso do primeiro programa não poriam seu trabalho em perigo, pois sua função seria a de lutar com o problema até achar algo que desse certo.

Simultaneamente, deveria ser instituída uma moratória para pesquisas de avaliação *ad hominem*, isto é, para pesquisas objetivando mais a avaliação de administradores específicos do que políticas administrativas. Se nos preocupamos com o problema do devassamento da intimidade nos bancos de dados e indicadores sociais do futuro (e.g., Sawyer e Schechter, 1968), o ponto mais inflamável seria o da intimidade dos administradores. Se o ameaçarmos, o sistema de medição será certamente sabotado por inúmeras formas possíveis. Embora isto possa parecer indevidamente pessimista, os casos freqüentes de administradores que tentaram arrasar achados de pesquisas indesejáveis convencem-me de que estou certo. Mas deveríamos poder avaliar as políticas alternativas que um dado administrador tem a opção de implementar.

2. EXPERIMENTOS DE CAMPO E ESQUEMAS QUASE-EXPERIMENTAIS

No esforço de estender a lógica da experimentação de laboratório para os trabalhos de "campo" e situações não-perfeitamente experimentais, organizamos uma lista de ameaças à validade experimental, em cujos termos cerca de 15 ou 20 esquemas experimentais e quase-experimentais foram avaliados (Campbell, 1957, 1963; Campbell e Stanley, 1963). Neste artigo, somente três ou quatro esquemas serão examinados e, portanto, nem todas as ameaças à validade serão relevantes, mas teremos um cenário útil para examiná-las sumariamente a todas. Seguem-se nove ameaças à validade interna.[1]

a) *História*: acontecimentos, que não o tratamento experimental, que ocorrem entre o teste prévio e o teste posterior, fornecendo assim uma explicação alternativa para os efeitos.

b) *Maturação*: processos internos dos respondedores ou das unidades sociais observadas, os quais produzem mudanças como resultado da passagem do tempo em si, tais como crescimento, fadiga, tendências seculares, etc.

c) *Instabilidade*: falta de confiabilidade das medidas, flutuações nas pessoas ou componentes que compõem a amostra, instabilidade autônoma de medidas repetidas ou "equivalentes". (Esta é a única ameaça para a qual os testes estatísticos de significância são relevantes.)

d) *Teste*: efeito da aplicação de um teste sobre o resultado de um segundo teste. Efeito da publicação de um indicador social sobre os valores subseqüentes daquele indicador.

e) *Instrumentação*: mudanças na calibração de um instrumento de medida ou mudanças nos observadores ou no sistema de medição, os quais podem ocasionar mudanças nas medidas obtidas.

f) *Ilusões criadas por regressão*: alterações falsas que ocorrem quando as pessoas ou unidades de tratamento

são selecionadas com base nos valores extremos de um seu atributo.

g) *Seleção:* vícios resultantes de um recrutamento diferencial dos grupos de comparação, ocasionando níveis médios diferentes na medida dos efeitos.

h) *Mortalidade experimental*: a perda diferencial de respondedores de grupos de comparação.

i) *Interação entre seleção e maturação*: vícios de seleção que ocasionam taxas diferentes de "maturação" ou mudança autônoma.

Se uma mudança ou diferença ocorre, estas são explicações rivais que poderiam ser usadas para explicar um efeito e assim negar que num experimento específico, qualquer efeito genuíno do tratamento experimental tenha sido demonstrado. São estas as falhas que os experimentos verdadeiros evitam, principalmente através do uso da escolha aleatória e dos grupos de controle. Na abordagem aqui preconizada, esta lista de verificação é usada para avaliar esquemas quase-experimentais específicos. Trata-se de uma avaliação, não de uma rejeição, pois acontece freqüentemente que, para um esquema específico, numa situação específica, a ameaça não é plausível, ou existem dados suplementares que possam ajudar a descartá-la mesmo quando a escolha aleatória é impossível. A ética geral, preconizada aqui tanto para administradores públicos como para cientistas sociais, é a de usar o melhor método possível, visando a experimentos verdadeiros com grupos de controle aleatórios. Mas quando o tratamento aleatório não é possível, preconiza-se o uso autocrítico de esquemas quase-experimentais. Precisamos fazer o melhor possível com o que nos é disponível.

Nossa posição face aos críticos perfeccionistas afeitos à experimentação de laboratório é mais militante do que isso: as únicas ameaças à validade que permitiremos invalidar um experimento são aquelas que reconhecem o *status* de leis empíricas que sejam mais confiáveis e mais plausíveis do que a lei que envolve o tratamento. A mera possibilidade de alguma explicação alternativa não é o bastante — somente as hipóteses rivais *plausíveis* as que são capazes de invalidar. Face aos estudos de correlação e estudos descritivos de bom-senso, por outro lado, nossa posição é a de uma maior cautela. Por exemplo, devido à armadilha metodológica específica da ilusão criada por regressão, a tradição sociológica dos esquemas *ex post facto* (Chapin, 1947; Greenwood, 1945) é totalmente rejeitada (Campbell e Stanley, 1963, p. 240-1; 1966, p. 70-1).

As ameaças à validade externa, focalizadas adiante, abrangem os problemas de validade enfrentados na interpretação dos resultados experimentais, as ameaças à generalização válida dos resultados para outras situações, para outras versões do tratamento, ou para outras medidas do efeito.[2]

a) *Efeitos de interação do teste*: o efeito de um teste prévio em aumentar ou diminuir a sensibilidade ou receptividade à variável experimental, fazendo, assim, os resultados obtidos para uma população previamente testada não-representativos dos efeitos da variável experimental para o universo não submetido ao teste prévio do qual os respondedores foram selecionados.

b) *Interação entre a seleção e o tratamento experimental*: sensibilidade não representativa da população tratada.

c) *Efeitos reativos dos preparativos para o experimento*: "artificialidade"; condições da situação experimental não são típicas das condições em que o tratamento é aplicado regularmente: "efeitos de Hawthorne".

d) *Interferência entre tratamentos múltiplos*: quando são aplicados conjuntamente múltiplos tratamentos, efeitos que não são típicos da aplicação separada dos tratamentos.

e) *Sensibilidade irrelevante das medidas*: todas as medidas são complexas e incluem componentes irrelevantes que podem ocasionar efeitos ilusórios.

f) *Repetibilidade irrelevante dos tratamentos*: os tratamentos são complexos e repetições dos mesmos podem deixar de incluir aqueles componentes que são na realidade responsáveis pelos efeitos.

Estas ameaças aplicam-se tanto aos experimentos verdadeiros como a quase-experimentos. São especialmente relevantes na experimentação aplicada. Na história cumulativa de nossa metodologia, este grupo de ameaças foi registrado pela primeira vez como crítica a experimentos verdadeiros que envolvam teste prévio (Schanck e Goodman, 1939; Solomon, 1949). Tais experimentos forneciam um fundamento legítimo para se generalizar a outras populações *previamente testadas*, mas as reações ao tratamento daquelas não-submetidas ao teste prévio poderiam ser bem diversas. Por essa razão preconiza-se experimentos verdadeiros, delineados de forma a dispensar o teste prévio (Campbell, 1957; Schanck e Goodman, 1939; Solomon, 1949) e uma busca de medidas não-reativas (Webb, Campbell, Schwartz e Sechrest, 1966).

Essas ameaças à validade servirão de base para a discussão que faremos de vários esquemas experimentais particularmente adequados à avaliação de programas específicos de melhoria social. Esses esquemas são os seguintes: "esquema da série temporal interrompida"; "esquema da série de controle"; "esquema da descontinuidade na regressão"; e vários "experimentos verdadeiros". A ordem que seguiremos é a dos esquemas fracos mas geralmente disponíveis para os mais fortes, que requerem mais previdência e determinação do administrador.

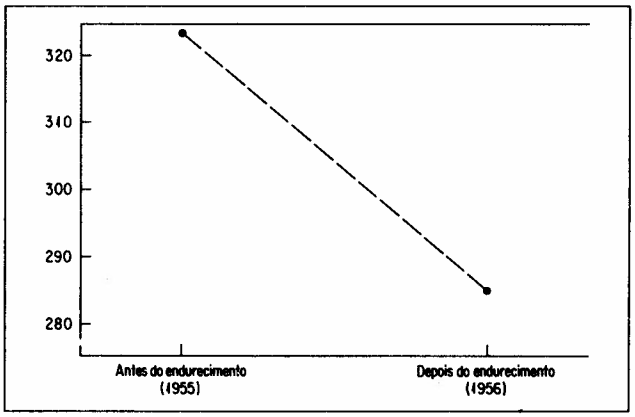
3. ESQUEMA DE SÉRIE TEMPORAL INTERROMPIDA

Normalmente quando uma unidade política inicia uma reforma, esta é instituída de modo geral, afetando toda a unidade. Nessa situação, a única base de comparação é

a documentação dos anos anteriores. A utilização usual é uma versão descuidada de um esquema quase-experimental muito fraco, o esquema de teste prévio e teste posterior de um só grupo.

Prova conveniente nos é dada pelo maior rigor adotado na punição do excesso de velocidade no Estado de Connecticut em 1955, analisado por mim e pelo sociólogo H. Laurence Ross como um exemplo para esclarecimento (Campbell e Ross, 1968; Glass, 1968; Ross e Campbell, 1968). Depois de se ter registrado o maior número já ocorrido de mortes em acidentes de trânsito em 1955, o Governador Abraham Ribicoff combateu com rigor sem precedentes o excesso de velocidade. Após um ano de execução das medidas adotadas ocorreram 284 mortes no trânsito, em comparação com 324 no ano anterior. Ao anunciá-lo, o governador declarou: "Tendo sido salvas 40 vidas em 1956, uma redução de 12,3% do número de mortos no trânsito em 1955, podemos reiterar definitivamente a validade do programa." Estes resultados estão no gráfico da figura 1, enfatizados deliberadamente para fazê-los parecer impressionantes.

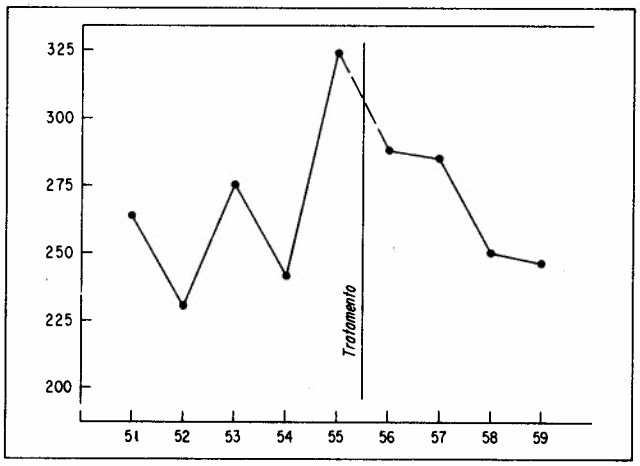
Figura 1 - Mortes ocorridas no trânsito no estado de Connecticut



No que se segue, embora reconheçamos que as medidas adotadas tiveram alguns efeitos benéficos, criticaremos a interpretação que Ribicoff deu aos seus resultados, do ponto de vista dos padrões estritos de evidência do cientista social. Não fosse o agora Senador Ribicoff homem da estatura que é, a crítica seria impolítica, porque estaríamos indispondo-nos com um dos mais fortes proponentes da experimentação social nos Estados Unidos. Devido à sua índole, no entanto, podemos sentir-nos seguros de que ele compartilha dos nossos interesses, tanto num programa progressivo de melhoria social experimental como na realização de avaliações mais sérias possíveis desses experimentos. Na verdade, foi sua integridade em usar todos os meios à sua disposição como governador para garantir que o impopular rigor contra o excesso de velocidade fosse de fato cumprido que torna esses dados dignos de qualquer exame. Mas as ricas possibilidades deste exemplo e nossa tentação política de substituí-lo por um outro que fosse menos melindroso demonstram os problemas políticos que precisam ser enfrentados quando se experimenta com reforma social.

Considerando a figura 1 e a declaração de Ribicoff, vamos observar os mesmos dados apresentados como parte de uma série temporal prolongada na figura 2, e examinar detalhadamente as ameaças relevantes à validade interna:

Figura 2 - Mortes ocorridas no trânsito em Connecticut. (Mesmos dados da figura 1 apresentados como parte de uma série temporal prolongada.)



a) *História*: as duas apresentações deixam de controlar os efeitos de outros agentes potenciais de mudanças. Por exemplo, 1956 pode ter sido um ano excepcionalmente seco, com menos acidentes causados pela chuva ou pela neve. Ou pode ter havido um acréscimo significativo no uso de cintos de segurança ou outras medidas de segurança. A estratégia que preconizamos na quase-experimentação não é a de erguer as mãos, num gesto de desistência, recusando o uso da evidência por falta do controle, mas sim a de gerar, através de crítica bem informada e apropriada para esta situação específica, tantas hipóteses rivais plausíveis quantas for possível e então fazer a pesquisa suplementar de, por exemplo, registros meteorológicos e de vendas de cintos de segurança, que poderiam afetar essas hipóteses rivais.

b) *Maturação*: este termo vem de críticas de estudos sobre treinamento de crianças. Aplicado aqui para os dados dos testes prévio e posterior da figura 1, a hipótese plausível poderia ser a de que as taxas de mortalidade estavam decrescendo de ano para ano (como de fato estão, nos Estados Unidos, em relação a milhas percorridas ou ao número de automóveis). Neste caso a série temporal prolongada apresenta grande vantagem

metodológica e descarta essa ameaça à validade. A tendência geral é inconsistentemente a de aumento antes do endurecimento e de uma diminuição estável depois.

c) *Instabilidade*: estava aparentemente implícito no pronunciamento oficial o pressuposto de que toda a alteração de 1955 e 1956 fora devida ao endurecimento. Não foi reconhecido o fato de que todas as séries temporais são instáveis mesmo quando nenhum tratamento é aplicado. O grau dessa instabilidade normal é a questão crucial, e uma das principais vantagens da série temporal prolongada é que ela apresenta uma amostra dessa instabilidade. A grande instabilidade anterior ao tratamento faz agora o efeito do tratamento parecer trivial. O salto de 1955-56 é menor do que os aumentos tanto de 1954-55 como de 1952-53. É verdade que é o maior decréscimo da série, mas supera os de 1951-52, 1953-54 e 1957-58 por valores triviais. Dessa forma, as instabilidades inexplicadas da série são tais que fazem com que o decréscimo de 1955-56 seja interpretável como uma variação como as demais. Por outro lado, deve ser notado que depois do endurecimento não houve mais aumentos, e, nesse sentido, a feição da série temporal parece indubitavelmente ter mudado.

A ameaça da instabilidade é a única para a qual os testes estatísticos de significância são relevantes. Box e Tiao (1965) têm um elegante modelo bayesiano para série temporal interrompida. Aplicado por Glass (1968) a dados mensais do nosso caso em foco — removidas as tendências sazonais — demonstra uma redução estatisticamente significante na série temporal após o endurecimento. Mas, como veremos, existe uma explicação alternativa para pelo menos parte desse efeito significante.

d) *Regressão*: nos experimentos verdadeiros o tratamento é aplicado independentemente do estado prévio das unidades. Em experimentos naturais o fato de um grupo ter sido submetido a tratamento é muitas vezes um dos sintomas de condição do grupo tratado. O tratamento pode então ser perfeitamente um *efeito* em vez de, ou além de, uma causa. A psicoterapia é um desses casos onde o tratamento é um dos sintomas, como o é qualquer caso em que o grupo tratado é auto-selecionado ou autodesignado por motivo de necessidade. Todos eles apresentam problemas especiais de interpretação, dos quais o exemplo presente é um tipo.

A hipótese rival plausível da seleção-regressão parte do seguinte argumento: dado que a taxa de mortalidade tem certo grau de variabilidade, então uma subamostra selecionada por causa do seu valor extremo em 1955 seria, em média, menos extrema em 1956, como mero reflexo dessa variabilidade. Houve seleção baseada em valor extremo na aplicação deste tratamento? Provavelmente sim. De todos os registros anuais de mortes cusadas pelo trânsito em Connecticut, a ocasião mais provável para um endurecimento com o excesso de velocidade seria após um ano de taxa excepcionalmente alta. Se a série temporal mostrava instabilidade, a taxa do ano seguinte seria em média menor, *unicamente em função dessa instabilidade*. Efeitos de regressão são provavelmente a forma que mais reaparece de ilusão de si mesmo na literatura de experimentação em reforma social. É difícil torná-los intuitivamente óbvios. Tentemos novamente. Tomemos qualquer série temporal que apresente variabilidade, mesmo se esta represente puro erro de medida. Percorramo-la como se seguíssemos o tempo. Escolhamos um ponto que é o "mais alto até então". Olhemos então o ponto seguinte. Na média dos casos, esse ponto será mais baixo, mais próximo da tendência geral.

Na situação que estamos examinando o salto mais notável em toda a série é o acréscimo imediatamente anterior ao endurecimento. É muito provável que esse aumento tenha originado o endurecimento, em vez de ter o endurecimento ocasionado a diminuição em 1956, ou pelo menos além de tê-la ocasionado. Pelo menos uma parte da queda em 1956 é um efeito do valor extremo de 1955. Embora o grau de regressão esperado possa em princípio ser computado a partir da autocorrelação da série, não temos nesse caso uma quantidade suficientemente extensa de dados para fazê-lo com alguma confiança.

O aconselhamento de administradores que queiram fazer testes genuínos da realidade deve dar atenção a este problema difícil de ser superado. O conselho mais geral é o de lidar com problemas crônicos cuja urgência ou cujos valores extremos sejam persistentes, em vez de reagir a um extremo momentâneo. O administrador deveria examinar a série temporal antes do tratamento para julgar se a instabilidade ou extremos momentâneos poderiam invalidar ou não os resultados do seu programa. Se o pudessem, deveria programar o tratamento para um ou dois anos mais tarde, de modo que sua decisão fosse mais independente do valor extremo daquele ano. (Os vícios de seleção que ainda permanecem neste procedimento precisam de um exame adicional.)

Ao dar conselhos ao administrador *experimental*, está-se dando inevitavelmente conselhos aos administradores *encurralados*, cuja embaraçosa situação política exige resultado favorável, seja válido ou não. A tais administradores encurralados, o conselho é o de escolher o pior ano de todos e a unidade social que seja de fato a pior. Se há instabilidade inerente, não há direção a seguir senão melhorar, ao menos em média.

Duas outras ameaças à validade interna merecem discussão com respeito a este esquema. Quando falamos em testar, ocorre-nos tipicamente a situação na qual um teste de atitude, aptidão ou personalidade é ele mesmo agente de mudança, ao persuadir, informar, treinar ou de qualquer outra forma, ao acionar processos de mudança. No caso que estamos analisando nenhum procedimento de teste foi introduzido artificialmente. Porém, para o simples esquema de "antes e depois" da figura 1, se o teste prévio fosse o primeiro dado sobre o assunto levantado e publicado, a publicidade, por si só, poderia ocasionar uma redução na taxa de mortes no trânsito, a qual teria tido lugar mesmo sem a adoção de medidas mais rigorosas contra o excesso de velocidade. Muitos programas de segurança no trânsito já pressupõem isso. A evidência fornecida por uma série temporal prolongada tranqüiliza-nos a esse respeito somente na medida em que podemos pressupor que os números foram publicados com ênfase equivalente em todos os anos.[3]

Mudanças de *instrumentação* não são uma falha provável neste exemplo, mas o seriam se tivesse havido uma alteração nos hábitos de registro ou na responsabilidade institucional ao mesmo tempo em que ocorreu o endurecimento. Num caso como este, provavelmente é melhor usar freqüências absolutas do que índices cujos parâmetros de correção estejam sujeitos a revisões periódicas. Por exemplo, taxas *per capita* estão sujeitas a saltos periódicos toda vez que os resultados de um recenseamento ficam disponíveis e as extrapolações feitas anteriormente são revistas. De forma análoga, uma mudança nos quilômetros por litro usado para se estimar a quilometragem total em taxas de mortalidade por quilômetro rodado poderia explicar uma variação nessas taxas. É claro que tais vícios podem também estar ocultando um efeito genuíno. É quase certo que o endurecimento de Ribicoff reduziu a velocidade no trânsito (Campbell e Ross, 1968). Um tal decréscimo em velocidade aumenta o rendimento de combustível em quilômetros por litro, de modo que se fosse usado o mesmo rendimento anterior para a estimativa de quilômetros rodados, como certamente o seria, obter-se-ia um valor subestimado e portanto um aumento ilusório na taxa de mortes por quilômetro rodado.

As reformas que introduzem modificações abruptas de política tendem também a modificar o sistema de registro de dados e assim confundir tratamentos da reforma com mudanças na instrumentação. O administrador experimental ideal fará o possível para evitá-lo. Preferirá manter um sistema de mensuração parcialmente imperfeito, mas comparável, a perder de vez a possibilidade de comparação. No entanto, a situação política torna às vezes isso impossível. Consideremos como uma reforma experimental a reorganização do sistema policial de Chicago feita por Orlando Wilson. A figura 3 mostra seu impacto em furtos de pouca monta em Chicago — um notável *aumento*! É claro que Wilson notou o impasse com antecedência, pois um dos aspectos da sua reforma foi a reorganização do sistema de registros. (Note-se nos dados anteriores à reforma, a ausência suspeita de uma tendência secular de aumento.) Nesta situação, Wilson não tinha outra alternativa. Se tivesse deixado o sistema de registros inalterado, com o intuito de obter um esquema experimental melhor, seus policiais ressentidos o teriam triturado com uma onda de crime, começando a registrar deliberadamente as muitas queixas que não vinham constando nos livros.[4]

Aqueles que advogam o uso de medidas baseadas em dados de arquivos como indicadores sociais (Bauer, 1966; Gross, 1966, 1967; Kaysen, 1967; Webb *et alii*, 1966) precisam enfrentar, sem medo, não só o seu alto nível de erro caótico e de vício sistemático como também as mudanças no sistema de registro motivadas por razões políticas que se seguem ao seu uso público como indicadores sociais (Etzioni e Lehman, 1967). As medidas não são igualmente suscetíveis. Na figura 4 o efeito de Orlando Wilson sobre os homicídios parece insignificante de uma forma ou de outra.

Das ameaças à validade externa, a mais relevante para a experimentação social é a *sensibilidade irrelevante das medidas*. Parece que é melhor discutir isso com respeito ou ao problema de se generalizar de um indicador para outro, ou à validade imperfeita de todas as medidas, que só é superada com o uso de medidas múltiplas, cujas imperfeições são independentes entre si (Campbell e Fiske, 1959; Webb *et alii*, 1966).

Para tratamentos de qualquer problema dado dentro de determinada subunidade governamental ou privada, haverá geralmente algo como um monopólio do governo em reforma. Mesmo que divisões diferentes estejam ten-

Figura 3 - Número registrado de furtos de pouca monta (menos de 50 dólares) em Chicago, de 1942 a 1962 (dados obtidos em Uniform crime reports for the United States, 1942-1962)

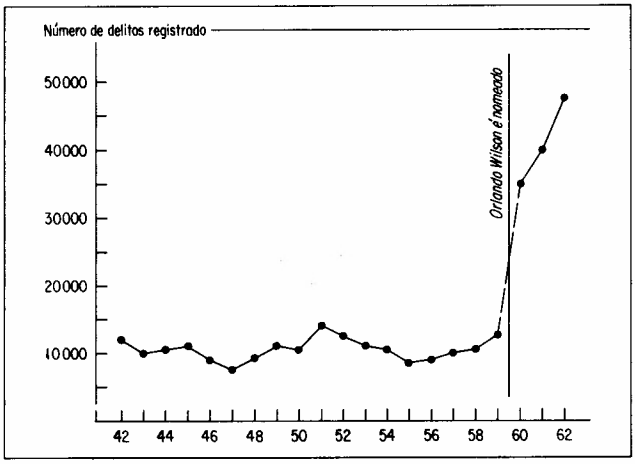
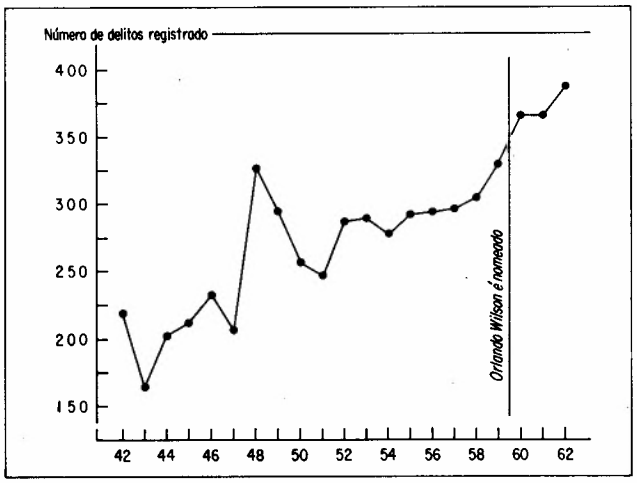


Figura 4 - Número registrado de homicídios dolosos e culposos em Chicago, de 1942 a 1962 (dados obtidos em Uniform crime reports for the United States, 1942-1962)



tando reformas diferentes da melhor forma possível, dentro de cada divisão geralmente haverá somente uma reforma em curso para um dado problema de cada vez. Mas para medidas de efeito isso não precisaria e nem deveria ser o caso. A própria máquina administrativa deveria propor medidas múltiplas de benefícios potenciais e de efeitos colaterais indesejáveis. Além disso, dever-se-ia permitir à oposição leal acrescentar ainda outros indicadores, com o processo político e o argumento adversário contestando tanto a validade como a importância relativa, com metodólogos das ciências sociais depondo para ambos os partidos e com os registros básicos mantidos públicos e sob auditoria bipartidária (como o são os votos eleitorais em condições ideais). Esse escrutínio competitivo é na verdade a principal fonte de objetividade nas ciências (Polanyi, 1966, 1967; Popper, 1963) e sintetiza um ideal de prática democrática em procedimentos tanto judiciais como legislativos.

As figuras seguintes retornam ao endurecimento com o excesso de velocidade em Connecticut e examinam outras medidas do efeito. São relevantes para confirmar que houve, de fato, um endurecimento e para a discussão dos efeitos colaterais. Também trazem o consolo metodológico de nos assegurar que em alguns casos o esquema da série temporal interrompida pode fornecer evidência clara de um efeito. A figura 5 mostra o salto na suspensão de carteiras de habilitação por excesso de velocidade — evidência de que uma punição severa foi instituída abruptamente. Mais um comentário para administradores experimentais: com este esquema fraco, *só mudanças abruptas e decisivas têm qualquer chance de ser avaliadas.* Uma reforma introduzida gradualmente será impossível de ser distinguida da circunstância de mudança secular, do efeito final de inúmeros agentes de mudança em ação contínua.

Gostaríamos de ter uma evidência intermediária de que a velocidade do trânsito foi alterada. Uma amostragem anual de algumas centenas de filmes de cinco minutos de cenas de auto-estradas (aleatória com relação ao local e à hora) poderia tê-la fornecido a custo moderado, mas os filmes não foram tomados. Dos registros públicos disponíveis talvez os dados da figura 6, que mostra a diminuição de multas por excesso de velocidade, indicam uma redução da velocidade do trânsito. Mas os efeitos do sistema legal eram complexos e em parte indesejáveis. O número de pessoas que guiavam com a carteira de habilitação cassada cresceu substancialmente (figura 7), pelo menos na amostra viciada dos que foram presos. Pode-se presumir que devido ao rigor da pena nos casos de culpa os juízes tenham-se tornado mais lenientes (figura 8), mas esse efeito é de significância marginal.

A relevância dos indicadores para os problemas sociais que queremos solucionar deve ser mantida constantemente em foco. A abordagem dos indicadores sociais tenderá a apontar como objetivo da ação social os próprios indicadores, em vez dos problemas sociais que eles indicam somente de forma imperfeita. Pode haver uma tendência de se legislar mudanças nos indicadores em si, em vez de mudanças nos problemas sociais.

Para exemplificar o problema da sensibilidade irrelevante das medidas, a figura 9 mostra um resultado da mudança na lei do divórcio efetuada na Alemanha em 1900. Numa reanálise recente dos dados, com a estatística de Box e Tiao (1965), Glass (Glass, Tiao e Maguire, 1969) concluiu que a mudança foi altamente significativa, ao contrário de análises estatísticas anteriores (Rheinstein, 1959; Wolf, Lüke e Hax, 1959). Mas a ênfase de Rheinstein ainda seria pertinente: a mudança nesse indicador não mostra melhora provável na harmonia e na estabilidade conjugais. Ao invés de reduzir a discórdia conjugal e a separação, a mudança legal tornou a taxa de divórcio um indicador menos válido desses fenômenos do que o era antes (ver também Etzioni e Lehman, 1967).

Figura 5-Cassações de cartas por excesso de velocidade, expressas em porcentagem de todas as cassações

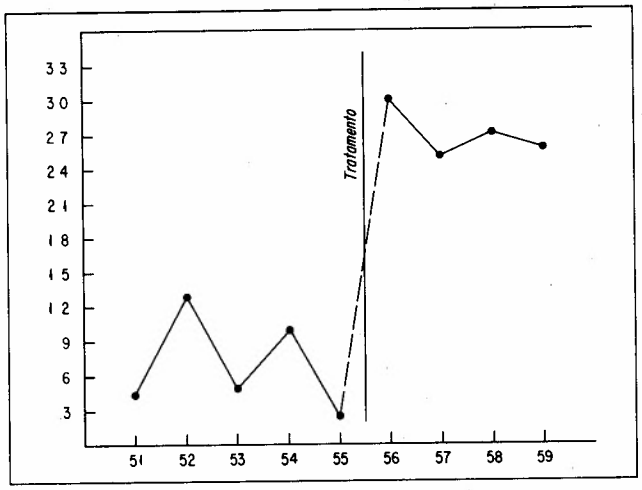
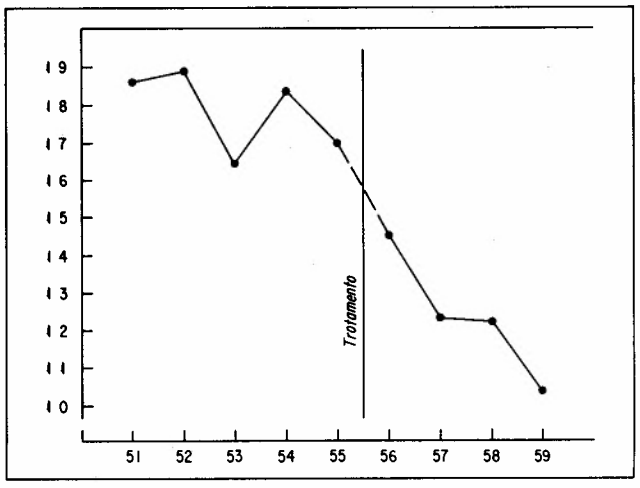


Figura 6-Multas por excesso de velocidade, expressas em porcentagem de todas as multas



4. ESQUEMA DA SÉRIE DE CONTROLE

O esquema da série temporal interrompida, como foi discutido até agora, pode ser usado em situações nas quais um grupo de controle é impossível, ou seja, naquelas em que a unidade governamental inteira recebeu o tratamento experimental constituído pela reforma social. No plano geral do delineamento quase-experimental, salientamos a grande vantagem de grupos de comparação não submetidos ao tratamento, mesmo quando estes grupos não podem ser designados aleatoriamente. O esquema mais comum desse tipo é o dos testes prévio e posterior com grupo de controle não equivalente, no qual, para cada um dentre dois grupos naturais, um deles recebe o tratamento, tomando-se duas medidas: uma num teste prévio e outra num posterior. Se evitarmos a prática tradicional, mas errônea, do emparelhamento baseado nos pontos obtidos no teste prévio (com os conseqüentes efeitos ilusórios causados por regressão), este esquema fornece um controle útil dos aspectos de história, maturação e efeitos de teste-reteste compartilhados pelos dois grupos. Mas não estabelece um controle para a hipótese rival plausível da *interação entre seleção e maturação*, isto é, a hipótese de que as diferenças de seleção nos agrupamentos naturais envolvem não só diferenças na média, mas também na velocidade de maturação.

Figura 7 Prisões de pessoas guiando com a carteira cassada, expressas em porcentagem sobre o total das cassações

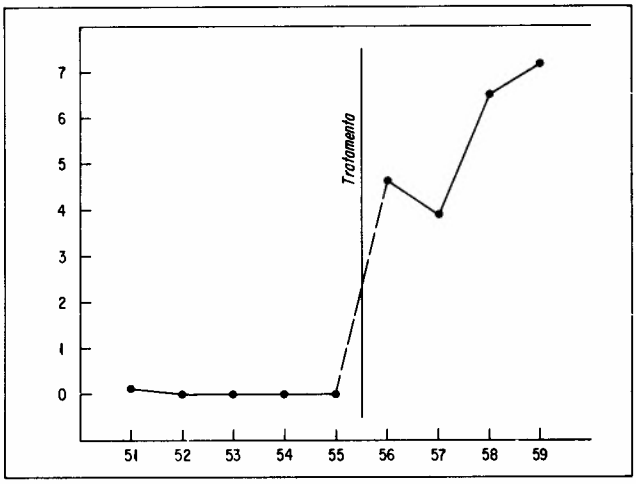


Figura 8-Porcentagem das multas por excesso de velocidade que foram depois canceladas por ter sido o acusado julgado inocente

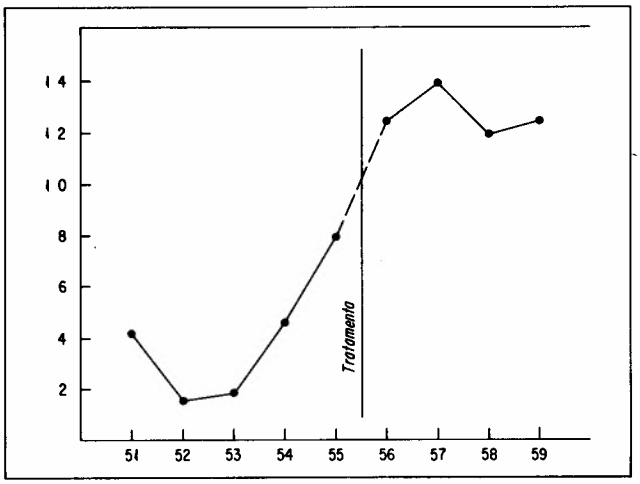


Figura 9-Taxa de divórcios do Império Alemão, 1881-1914

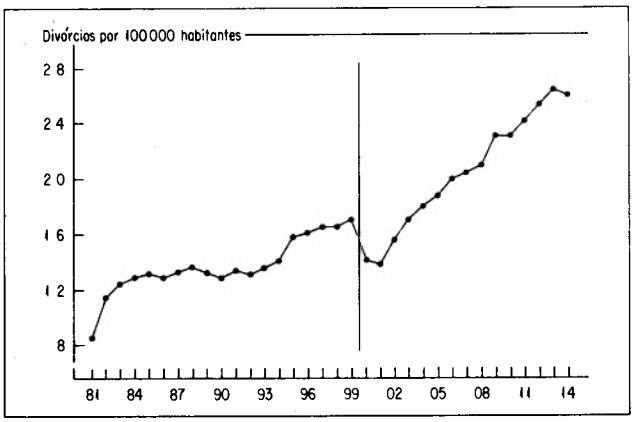
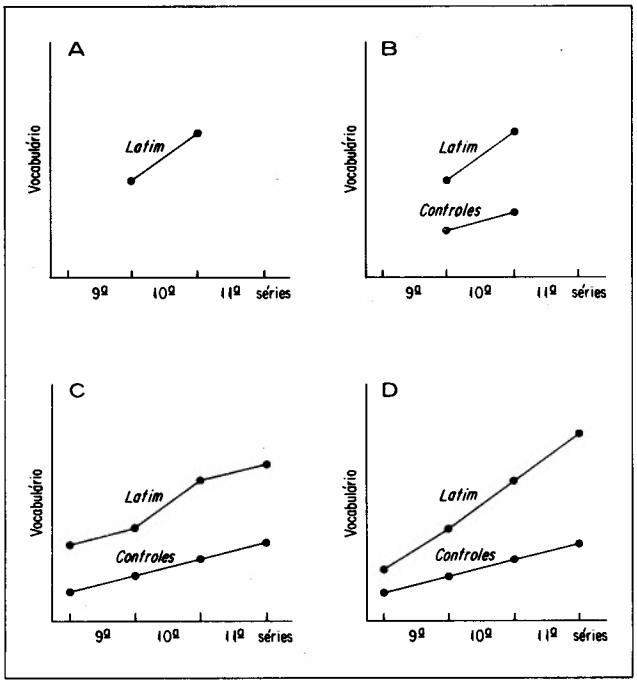


Figura 10 - Formas de análise quase-experimental do efeito de uma disciplina específica, incluindo-se o esquema da série de controle.



Esse argumento pode ser exemplificado com o problema de esquema tradicional quase-experimental dos efeitos do aprendizado do latim na aquisição do vocabulário em inglês, para estudantes americanos (Campbell, 1963). Nos dados hipotéticos da figura 10B, duas interpretações alternativas são possíveis. O latim pode ter causado um efeito, pois os que o estudaram lucraram mais do que os outros. Mas, por outro lado, os estudantes que querem aprender latim podem ter um aumento de vocabulário maior que se manifestaria mesmo que não tivessem estudado latim. Ampliando este esquema comum para duas séries temporais, obtemos uma evidência relevante, como demonstra a comparação dos dois resultados alternativos das figuras 10C e 10D. Dessa forma, aproximando-nos de um esquema quase-experimental, seja melhorando o esquema do grupo de controle não equivalente, seja melhorando o esquema da série temporal interrompida, chegaremos ao esquema da série de controle. A figura 11 mostra este esquema para o endurecimento com o excesso de velocidade em Connecticut, acrescentando evidência com as taxas de mortalidade de estados vizinhos. Aqui, os dados são apresentados na forma de taxas de mortalidade baseadas na população, para tornar as duas séries de magnitude comparável.

Figura 11-Esquema da série de controle, comparando o número de mortes em Connecticut com os de outros quatro estados

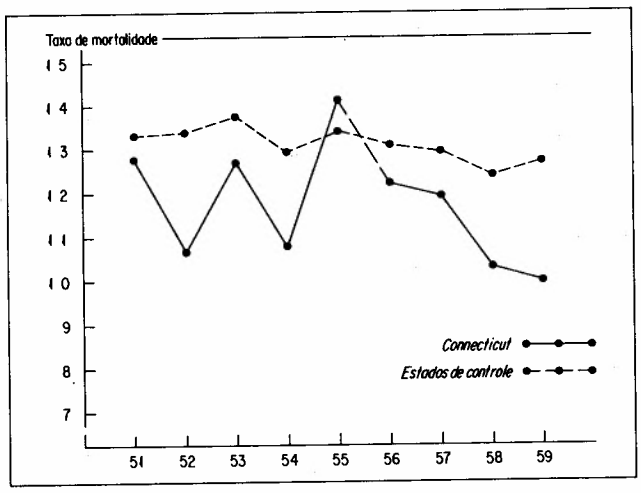
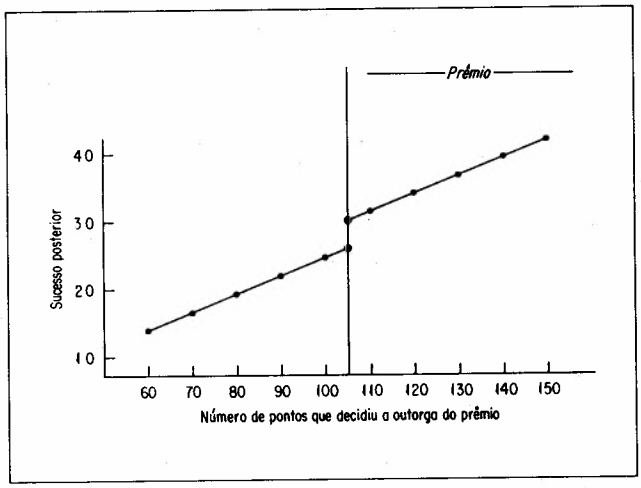


Figura 12 - Experimento de desempate e análise de descontinuidade na regressão



O esquema da série de controle da figura 11 mostra que havia uma tendência de decréscimo nos outros estados em 1955-56, devido provavelmente à história e maturação, isto é, a tendências seculares comuns, condições climáticas, dispositivos de segurança nos automóveis, etc. Mas os dados também mostram uma tendência de a taxa de mortalidade em Connecticut aproximar-se da dos outros estados antes de 1955 e de decrescer mais depressa que a dos outros estados a partir de 1956. Glass (1968) utilizou os dados mensais de Connecticut e dos estados de controle para gerar uma diferença mensal, a qual também mostra uma mudança significativa da tendência com a estatística de Box e Tiao (1965). Impressionados especialmente com a tendência de 1957, 1958 e 1959, estaremos dispostos a concluir que o endurecimento teve algum efeito além dos inegáveis pseudo-efeitos de regressão (Campbell e Ross, 1968).

As vantagens do esquema da série de controle demonstram os proveitos que a experimentação social pode tirar de um sistema social que permita diversidade nas subunidades. A possibilidade que tivemos de estimar os efeitos do endurecimento com o excesso de velocidade, assim como a que tiveram Rose (1952) e Stieber (1949) de estimar os efeitos das leis de arbitragem compulsória sobre as greves e a de Simon (1966) estimando a elasticidade-preço das bebidas alcoólicas, devem-se todas ao fato de que as mudanças não foram postas em vigor em todos os estados simultaneamente, por serem questões de alçada estadual e não federal. Embora não estejamos justificando desta forma uma diversidade desperdiçadora e injusta de leis e praxes de imposição de estado para estado, recomendaríamos enfaticamente que os engenheiros sociais fizessem uso dessa diversidade enquanto ela permanece disponível e que planejassem cooperativamente suas mudanças em política administrativa e no sistema de registro de modo a permitir uma inferência experimental ótima. Mais importante é a recomendação de que, para os aspectos de reforma social tratados pelo Governo federal, fosse considerada uma diversidade propositada na implementação, de modo a tornar disponíveis grupos de controle para análise. Se planejados corretamente, esses experimentos podem aproximar-se de experimentos verdadeiros, melhores do que os grupos de comparação fortuitos e *ad hoc* de que dispomos agora. Mas sem tal planejamento

fundamental, um controle central uniforme pode reduzir as possibilidades atuais de teste da realidade, ou seja, de uma experimentação social verdadeira. Dentro do mesmo espírito, a descentralização das tomadas de decisão, tanto dentro do Governo como dentro de monopólios privados, pode proporcionar uma concorrência útil à eficiência e à inovação manifestada numa multiplicidade de indicadores.

5. ESQUEMA DA DESCONTINUIDADE NA REGRESSÃO

Passaremos a considerar, agora, melhorias sociais que são escassas e que, portanto, não podem ser estendidas a todos os indivíduos. Essa escassez é inevitável em muitas circunstâncias e pode tornar possível uma previsão dos efeitos que do contrário seria impossível. Consideremos os notáveis experimentos da vacina Salk para poliomielite, nos quais ministrava-se a vacina a algumas crianças, enquanto que a outras aplicava-se uma injeção de um placebo salino inerte. Muitas dessas crianças do grupo de controle, atacadas mais tarde pela doença não teriam morrido se houvessem tomado a vacina real em lugar da droga inerte. A criação desses grupos de controle submetidos à pseudovacina teria sido impossível do ponto de vista moral, psicológico e social se tivesse havido vacina real para todos. Na ocasião, devido à escassez da vacina, a maior parte das crianças ficaria sem ela de qualquer forma. A criação dos grupos experimental e de controle foi uma forma altamente moral de distribuição daquela escassez, de modo a nos permitir o conhecimento da eficácia real do suposto bem. A prática médica usual de introduzir novas curas, experimentando-as na clínica geral, torna impossível uma avaliação, por confundir o estado prévio com o tratamento, isto é, por ministrar a droga aos mais necessitados ou mais desesperançados. Apresenta ainda o vício social de ministrar o suposto benefício aos membros das classes média e alta, mais assíduos em levar suas necessidades médicas ao conhecimento da comunidade médica. A postura política que favorece a experimentação social neste caso é o reconhecimento da distribuição aleatória como o meio mais democrático e moral de se alocar recursos limitados (e raros deveres arriscados), além do imperativo moral de utilizar essa distribuição aleatória de forma que a sociedade possa realmente conhecer o verdadeiro valor do suposto benefício. Esta é a ideologia que torna possível a realização de "experimentos verdadeiros" num grande número de reformas sociais.

Mas se a distribuição aleatória não for politicamente viável ou moralmente justificável numa dada situação, existe um poderoso esquema quase-experimental que permite que o bem escasso seja dado aos mais necessitados ou mais merecedores. Trata-se do esquema da descontinuidade na regressão que exige tão-somente uma atenção rigorosa e ordenada à dimensão da prioridade. Ele teve sua origem na defesa de um experimento de desempate na medida dos efeitos do recebimento de uma bolsa de estudos (Thistlethwaite e Campbell, 1960), e parece mais fácil explicá-lo à luz daquele experimento. Consideremos, como na figura 12, a dimensão aptidão e mérito antes do prêmio, a qual teria alguma relação com o sucesso posterior na vida (obter diploma em faculdade, o salário de 10 anos mais tarde, etc.). Aqueles que tinham a medida mais alta antes do prêmio são os mais merecedores e recebem o prêmio. Eles saem-se melhor na vida, mas será que o prêmio exerce alguma influência? Normalmente é impraticável dizê-lo, porque eles teriam se saído melhor de qualquer forma. Uma distribuição totalmente aleatória do prêmio teria sido impossível dada a intenção declarada de premiar o mérito e a aptidão. Mas seria viável tomar uma estreita faixa de aptidão em torno do valor que determinava quem receberia ou não o prêmio. Essas pessoas seriam consideradas como empatadas e dar-se-ia o prêmio a metade delas, através de escolha aleatória de desempate.

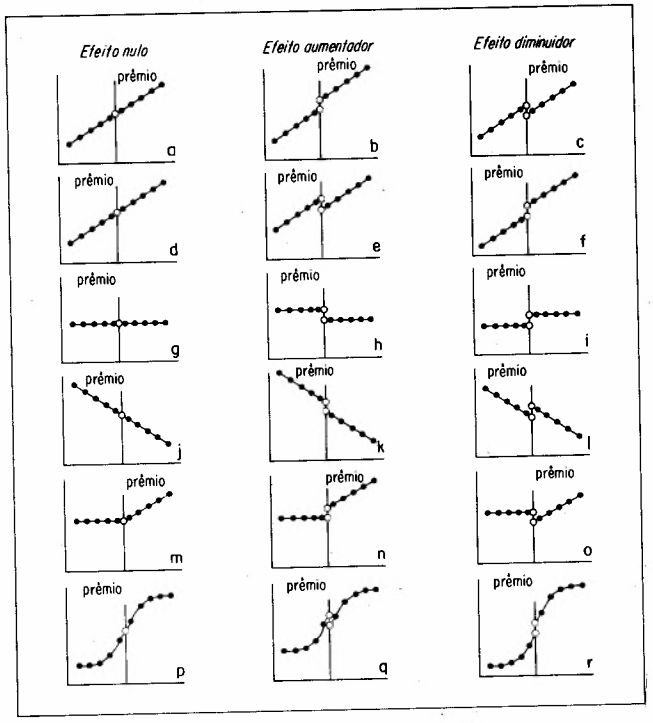
O fundamento lógico do experimento de desempate ainda o torna digno de ser realizado, mas ao considerar-se esse esquema ficou óbvio que se a regressão da medida antes do prêmio sobre efeitos posteriores fosse razoavelmente sistemática, poder-se-ia extrapolar os resultados do experimento de desempate, construindo dois gráficos da regressão do teste posterior (êxito após o prêmio) sobre o prévio (número de pontos baseados nos quais o prêmio foi dado), um para os que estavam na região dos premiados e outro para os situados na dos não-premiados. Se não houver diferença significativa para aqueles que estão na interseção das linhas de regressão com a linha de separação, então o experimento de desempate não deveria apresentar nenhuma diferença. Nos casos em que os que foram desempatados aleatoriamente mostrassem mais tarde um efeito como conseqüência de terem recebido o prêmio, deveria haver uma descontinuidade abrupta na linha de regressão. Tal descontinuidade não pode ser invalidada pela regressão normal que deve existir entre o teste posterior e o prévio, pois essa regressão normal, baseada numa amostra extensa das áreas de premiados e não-premiados, não justifica essa expectativa.

A figura 12 apresenta um exemplo no qual um número de pontos mais alto no teste prévio teria levado a um número de pontos mais alto no teste posterior, mesmo sem o tratamento, mas no qual há, além disso, um efeito substancial do tratamento. A figura 13 mostra uma série de resultados emparelhados, interpretando-se os da esquerda como não mostrando nenhum efeito e os da direita como mostrando um efeito. Note-se alguns casos peculiares. Em casos em que é dada uma oportunidade com base no mérito, como 13a e 13b (e a figura 12), um esquecimento da regressão subjacente do teste posterior sobre o teste prévio leva a pseudo-efeitos otimistas: na figura 13a, os que recebem o prêmio realmente saem-se melhor na vida, embora não seja, na verdade, por causa do prêmio. Mas em casos em que se procura incentivar os menos dotados, a situação tende a ser a das figuras 13d e 13e, em que o esquecimento da regressão subjacente fará o programa parecer nocivo se não houver efeito real, ou ineficaz se o houver.

É claro que o esquema funcionará igualmente bem ou até melhor se a dimensão que decide a outorga do prêmio — a medida do teste prévio — não tiver relação com a dimensão do teste posterior, ou se for irrelevante ou injusta, como nas figuras 13g, 13h e 13i. Em tais casos, a decisão da outorga do prêmio tem o mesmo efeito de uma distribuição aleatória. Relações subjacen-

tes negativas são obviamente possíveis, como nas figuras 13j, 13k e 13l. As figuras 13m, 13n e 13o foram incluídas para enfatizar que é o salto na interseção com o ponto de separação que demonstra o efeito, e que diferenças em inclinação que não sejam acompanhadas de diferenças no ponto de separação não são aceitáveis como evidências de efeito. Isto fica mais óbvio se lembrarmos que em casos como 13m, uma escolha aleatória de desempate não teria demonstrado diferença alguma. Relações subjacentes curvilíneas, como as das figuras 13p, 13q e 13r, criarão obstáculos adicionais à inferência clara em muitos casos em que o erro de amostragem poderia fazer com que a figura 13p se parecesse com a figura 13b.

Figura 13 - Exemplos de resultados de análises de descontinuidade na regressão.



Como exemplo adicional, a figura 14 apresenta dados simulados em computador, mostrando observações individuais e retas de regressão ajustadas a elas, numa versão mais completa do resultado de ausência de efeito da figura 13a. A figura 15 mostra um resultado com efeito. Esses dados foram gerados[5] atribuindo-se a cada indivíduo um número aleatório ponderado da distribuição normal como um "número de pontos verdadeiro", ao qual é adicionado um "erro" independente, também ponderado, para se obter a medida do teste prévio. O "número de pontos verdadeiro" adicionado a um outro "erro", gerado de forma análoga, produz a medida do teste posterior em casos de ausência de efeito, como o da figura 14. Na simulação de presença de efeito, como o da figura 15, são adicionados "pontos refletindo o efeito" à medida do teste posterior de todos os casos que receberam o tratamento, ou seja, aqueles além do ponto de separação na medida do teste prévio.[6]

Figura 14 - Esquema de descontinuidade na regressão: Efeito nulo

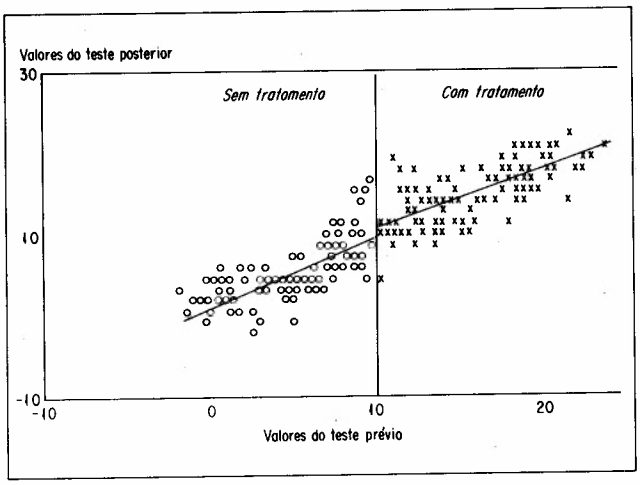
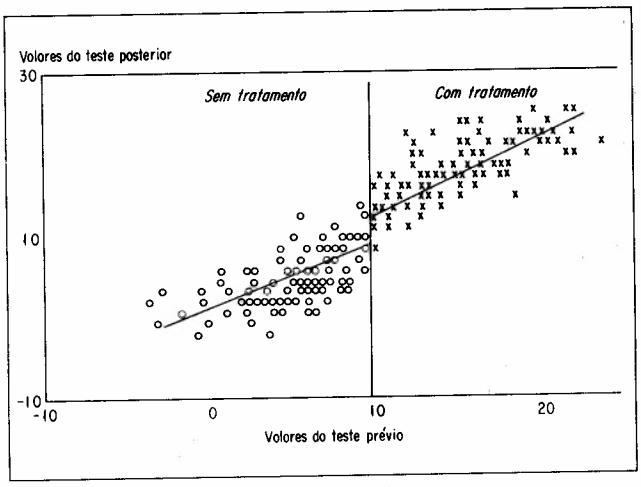


Figura 15 - Esquema de descontinuidade na regressão: Efeito autêntico



Este esquema poderia ser utilizado em várias situações. Consideremos os inscritos para o Corpo de Treinamento para Empregos (Job Training Corps), em maior número do que o programa pode atender, com a qualificação determinada pela necessidade. A situação seria a das figuras 13d e 13e. A dimensão básica para a decisão poderia ser a renda familiar *per capita*, sendo que aqueles que tivessem menos que um certo valor receberiam o treinamento. A dimensão para se medir o resultado do programa poderia ser o imposto de renda

retido na fonte dois anos mais tarde, ou percentagem dos que recebem seguro de desemprego. Tais valores de acompanhamento seriam fornecidos pelo Banco Nacional de Dados através do número de inscrição na previdência social, sem quebrar o anonimato individual e o sigilo pessoal, pois é o programa que está sendo examinado, através de dados agregados de muitas pessoas. Embora se pudesse dar nomes aos pontos individuais, isso não é necessário. Num clássico experimento de campo sobre obediência ao pagamento de impostos, Richard Schwartz e o Bureau of Internal Revenue (equivalente à nossa Secretaria da Receita Federal) conseguiram juntar grupos de entrevistas pessoais e declarações de imposto de renda de modo a permitir análises estatísticas sem que os diferentes encarregados, tanto das entrevistas como das declarações, ficassem sabendo os dados correspondentes de nenhuma pessoa especifica (Schwartz e Orleans, 1967; ver também Schwartz e Skolnick, 1963). Manniche e Hayes (1957) já explicaram detalhadamente como se pode usar um intermediário para emparelhamento em dois estágios de dados duplamente codificados. Kaysen (1967) e Sawyer e Schechter (1968) apresentam discussões sensatas do problema mais geral.

O que se requer do administrador de um bem melhorador escasso para que se utilize esse esquema? O mais essencial é um nítido ponto de separação ao longo da dimensão que constitui o critério de decisão e ao longo da qual outros pontos de separação possam ser analogamente escolhidos, tanto acima como abaixo do ponto de separação utilizado para o prêmio. Isso ficará mais claro mostrando-se por que a entidade que concede as bolsas do Mérito Nacional não pode usar o esquema para a real decisão da concessão (embora o tivesse usado para o Certificado de Mérito). No seu sistema de trabalho, cada uma dentre várias comissões toma decisões de concessão do prêmio considerando um grupo de candidatos e escolhendo os *N* melhores para receber as *N* bolsas disponíveis. Esse procedimento fornece um ponto de separação ao longo de uma dimensão não-especificada que é uma mistura de critérios, mas deixa de fornecer pontos potenciais de separação acima e abaixo. O que poderia ser feito é que cada comissão classificasse, de forma coletiva, o seu grupo de candidatos, que são em torno de 20. Os *N* melhores receberiam, então, o prêmio. Ao combinar os casos das várias comissões cada caso poderia ser classificado de acordo com sua posição em relação ao ponto de separação que decidiu o prêmio, fosse acima ou abaixo deste. Para efeito da regressão com as medidas pós-tratamento, essa classificação seria análoga ao ponto de separação. Tal classificação dos grupos consumiria tempo das comissões. Procedimento igualmente aceitável, se as comissões concordassem, seria o de fazer cada membro da comissão atribuir a cada candidato uma nota, *A+*, *A*, *A-*, *B+*, *B*, etc., após ampla discussão com liberdade de revisão, e conceder a bolsa aos *N* candidatos que obtivessem a melhor média nessa avaliação, não sendo permitidas revisões após o cômputo das médias. Essas unidades de classificação, mesmo que não fossem comparáveis entre uma comissão e outra na faixa de talento abrangida, no número de pessoas classificadas ou no valor que serviu de ponto de separação, poderiam ser combinadas sem vício para se analisar a descontinuidade na regressão, na região de valores acima e abaixo do ponto de separação em que todas as comissões estivessem representadas.

É a dimensionalidade e a nitidez do critério de decisão que está em debate, não os seus componentes ou sua validade. As classificações poderiam ser feitas na base de nepotismo, capricho e superstição e, mesmo assim, servirem. Como já foi dito, se o critério de decisão é completamente inválido, aproximamo-nos da distribuição aleatória dos experimentos verdadeiros. Portanto, a fraqueza das decisões subjetivas das comissões não é sua subjetividade, mas sim o fato de que elas fornecem somente um ponto de separação na sua dimensão subjetiva final. Os procedimentos recomendados, mesmo na forma de notas médias, provavelmente representam pequeno aumento na carga de trabalho das comissões. Mas isso poderia ser justificado perante essas comissões lembrando-lhes que, devido às desistências, etc. não se sabe exatamente o número de pessoas a quem serão concedidas bolsas quando a comissão se reúne. Outros custos na ocasião do planejamento são igualmente mínimos. A sobrecarga principal é manter bons registros tanto dos que receberam o prêmio como dos que não o receberam. Dessa forma, um administrador experimental pode, a um custo baixo, lançar os fundamentos para um acompanhamento científico posterior, para os quais nem se cogita ainda fazer orçamentos.

A situação que estamos analisando tende a ser mais uma onde as medidas de pré-tratamento, medidas de aptidão, avaliação das referências, etc., podem ser combinadas mediante correlação múltipla num índice único, que apresenta uma alta correlação, porém não perfeita com a decisão da concessão da bolsa. Se esse índice for usado como dimensão de teste prévio para a análise da descontinuidade na regressão, haverá então um ponto de separação indistinto. O esquema pode ser usado nesse caso? Provavelmente não. A figura 16 mostra o pseudo-efeito possível se a decisão da concessão contribui com qualquer variância válida para a evidência quantificada do teste prévio, como em geral é o caso. A reta de regressão do grupo premiado está acima da do grupo não-premiado somente por causa da variância válida neste caso simulado, não havendo nenhum efeito genuíno do prêmio. (Na simulação desse caso, a decisão da concessão da bolsa foi baseada num valor composto da medida verdadeira do teste prévio e de um erro independente.) A figura 17 mostra um ponto de separação indistinto mas com um efeito genuíno do prêmio.[7] A recomendação para o administrador fica clara: procurar estabelecer um ponto de separação nítido ao longo de um critério de decisão quantificado. Se existirem regras complexas de seleção, das quais só uma é quantificada, procurar fazer um acompanhamento do subconjunto das pessoas para as quais a dimensão quantificada foi decisiva. Se um *pistolão* político criar algumas decisões inconsistentes com o ponto de separação, registrar esses casos como baseados numa "regra de decisão qualitativa" e mantê-los fora da sua análise experimental.

Quase todos nossos programas de melhorias planejados para os menos privilegiados poderiam ser estudados por meio deste esquema, assim como algumas ações importantes do Governo que afetam as vidas dos cidadãos de formas que não julgamos ser experimentais. Por exemplo, durante um período considerável o nú-

mero de pontos obtidos em testes tem sido usado na convocação para o serviço militar ou para rejeitar como incapaz na faixa mais baixa de aptidão. Se esses pontos de separação, número de pontos obtidos nos testes, nomes e números de previdência social foram registrados para alguns intervalos acima e abaixo do ponto de separação, poderíamos fazer estudos elegantes do efeito do serviço militar na renda posterior, mortalidade, número de dependentes, etc. Infelizmente para esse objetivo, a operação conhecida como "Operation 100,000", instituída pelo Secretário da Defesa com nobres intuitos experimentais, está tornando indistinto o ponto de separação. Mas dispomos de dados anteriores a vários anos, referentes ao Vietnã, prontos para análise.

Figura 16 - Esquema de descontinuidade na regressão: Ponto de separação indistinto, com pseudo-efeito do tratamento

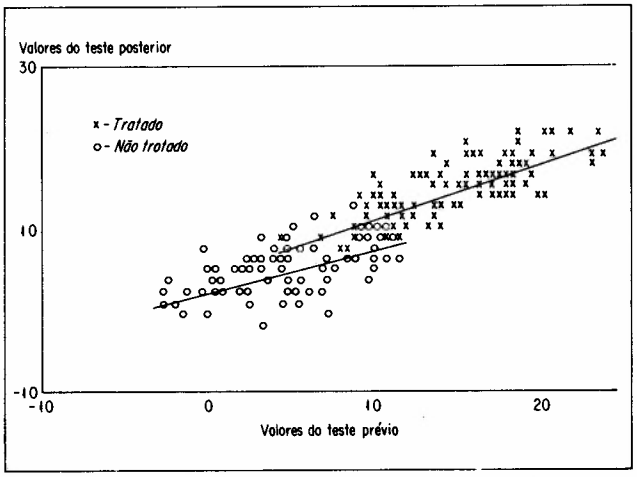
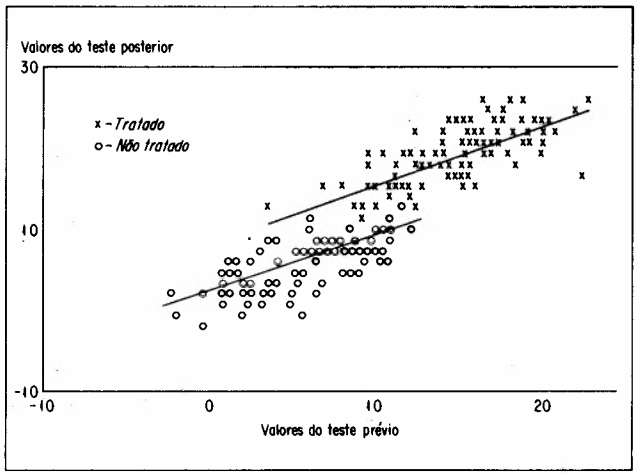


Figura 17 - Esquema de descontinuidade na regressão: Ponto de separação indistinto, com pseudo-efeitos somados a um efeito real do tratamento



Esse exemplo chama a atenção para uma das ameaças à validade externa desse esquema ou do experimento de desempate. O efeito do tratamento foi estudado somente para aquela estreita faixa de talento em torno do ponto de separação. Uma generalização dos efeitos do serviço militar, por exemplo, sobre a carreira dos mais aptos, feita com base num nível de aptidão muito baixo, seria extremamente arriscada. Mas nas leis de alistamento e nos requisitos do serviço militar pode haver outros pontos de separação nítidos ao longo de um critério quantitativo que também poderiam ser usados. Por exemplo, os que têm mais de seis pés e seis polegadas (1,98m) de altura são dispensados do serviço militar. Imagine um acompanhamento feito cinco anos mais tarde dos convocados agrupados por polegadas na faixa de seis pés e uma polegada a seis pés e cinco polegadas e de um grupo de seus correspondentes que teriam sido convocados não fosse a sua altura excessiva, seis pés e seis polegadas a seis pés e 10 polegadas. (A possibilidade de que outras razões para dispensa não terem sido examinadas pela junta de alistamento poderia ser um problema nesse caso mas, provavelmente, não insuperável.) O fato de que não se deveria esperar que a altura nessa faixa tivesse qualquer relação com variáveis na vida subseqüente não é absolutamente uma fraqueza desse esquema e se de fato tivermos uma subpopulação para a qual há um ponto de separação numérico nítido, conseguiremos obter uma medida de efeitos com validade interna. A dispensa no sistema atual é uma decisão não-quantificada de uma comissão. Mas, assim como o senso de justiça dos soldados americanos foi quantificado por meio da comparação de pares de casos de modo a se criar um sistema aceitável de pontos para baixa ao fim da II Guerra Mundial (Guttman, 1946; Stoffer, 1949), igualmente poderíamos conseguir quantificar um índice composto de prioridade para dispensa e aplicá-lo como critério uniforme em todo o país, estabelecendo-se assim outro ponto de separação numérico.

Além dos indicadores do tipo fornecido pelo Banco Nacional de Dados, haverá ocasiões em que serão necessárias novas coletas de dados através de entrevistas ou questionários, surgindo, então, o problema especial de cooperação desigual que poderia ser classificada como erro instrumental. No nosso modo tradicional de pensar, a perfeição da descrição é considerada mais valiosa do que a comparabilidade. Portanto, se como no estudo das bolsas, um questionário de acompanhamento enviado pelo órgão que as distribui apresentasse maior retorno dos que ganharam a bolsa, isso poderia parecer desejável, mesmo se o retorno das respostas dos que não a ganharam fosse muito menor. Do ponto de vista da quase-experimentação, no entanto, seria melhor usar uma agência de pesquisa independente e um objetivo dissimulado, obtendo-se assim taxas de resposta igualmente baixas, tanto dos que ganharam a bolsa como dos que não a ganharam e evitando-se a descontinuidade no grau de cooperação, a qual poderia ser interpretada erroneamente como uma descontinuidade em efeitos mais importantes.

6. EXPERIMENTOS COM GRUPOS DE CONTROLE DESIGNADOS ALEATORIAMENTE

Experimentos com aleatorização (designação aleatória para o tratamento) tendem a ser limitados ao laboratório e ao posto de experimentos agrícolas. Mas certamente não é necessário que assim seja. A unidade de aleatorização pode ser pessoas, famílias, zonas eleitorais ou unidades administrativas maiores. Para objetivos estatísticos, as unidades de aleatorização devem ser numerosas e, portanto, teoricamente pequenas. Mas por razões de validade externa, inclusive preparativos reativos, as unidades de aleatorização deveriam ser escolhidas com base nas unidades de acesso administrativo. Quando as diretrizes são aplicadas através de contatos individuais com os clientes, pode-se conseguir uma aleatorização ao nível pessoal que não chame atenção, já que os clientes não precisam ficar sabendo que alguns deles recebem o tratamento e outros não. Todavia para a maioria das reformas sociais, unidades administrativas maiores estarão envolvidas, tais como salas de aula, escolas, cidades, municípios ou estados. Temos que desenvolver posturas e ideologias políticas que tornem possível a aleatorização a esses níveis.

"Projeto-piloto" é um termo útil que já consta do nosso vocabulário político. Significa um programa tentativo que, se funcionar, será ampliado para outras áreas. Modificando-se a prática real a esse respeito, sem sair do entendimento popular do termo, poder-se-ia desenvolver uma valiosa ideologia experimental. Como se escolhe uma área para projeto-piloto? Se o público estiver preocupado com isso, a escolha provavelmente toma a forma de pressões junto aos legisladores, as quais representam somente em parte a maior necessidade de uma região, pois o poder e a conveniência políticos têm um papel importante. Sem violar a tolerância ou o propósito do público, poder-se-ia provavelmente arquitetar um sistema no qual as pressões sobre legisladores decidissem quais as áreas elegíveis para participarem de um sorteio público formal que realizaria as escolhas definitivas entre parelhas. Tais procedimentos de decisão, como tirar a sorte, são justamente respeitados já há muito tempo (por exemplo, Aubert, 1959). Atualmente mantêm-se registros nos projetos-pilotos somente para o grupo experimental, na maioria dos casos. De acordo com a ideologia experimental, seriam coletados dados comparáveis de controles designados. (É claro que há exceções da prática usual, como nos diligentes experimentos sobre os efeitos do flúor conduzidos pelo Serviço de Saúde Pública, nos quais, ano após ano, foram examinados os dentes de crianças de Oak Park, servindo de controle para aquelas tratadas em Evanston.) (Blayney e Hill, 1967.)

Outra postura política que torna possível a melhoria social experimental é a da *inovação gradativa*. Mesmo que a intenção seja a de implantar a reforma em todas as unidades, a logística da situação geralmente mostrará que uma introdução simultânea não é possível. O resultado é uma seqüência de conveniência a esmo. Num programa de inovação gradativa, a introdução do programa seria deliberadamente ampliada e as unidades escolhidas para serem as primeiras ou as últimas designadas por sorteio (talvez num sorteio entre parelhas de unidades), de forma que durante o período de transição os primeiros recipientes pudessem ser analisados como unidades experimentais e os últimos, como controles. Uma terceira ideologia que torna possível a realização de experimentos verdadeiros já foi discutida: a aleatorização como uma forma democrática de se distribuir recursos escassos.

Neste artigo não dedicaremos tanto espaço à experimentação verdadeira quanto à quase-experimentação, em virtude de existirem à nossa disposição discussões excelentes e fontes de consulta estatística para experimentos verdadeiros. Quando se pode fazer tanto experimentos verdadeiros como quase-experimentos, os primeiros devem ser quase sempre preferidos. Só ocasionalmente existem ameaças tão fortes à validade externa no experimento verdadeiro que um quase-experimento seria preferível. A distribuição de espaço neste artigo não deve ser interpretada de outra forma.

7. MAIS CONSELHOS PARA ADMINISTRADORES ESCURRALADOS

Na realidade, a rivalidade não se dá entre os quase-experimentos aqui revistos, os quais são razoavelmente interpretáveis, e os experimentos "verdadeiros". Ambos representam raras eminências em comparação com uma visão distorcida e enganosa de si mesmo. Tanto para enfatizar esse contraste, como para sugerir novamente uma orientação que beneficie aos administradores encurralados, cuja embaraçosa situação política não permitirá o risco do fracasso, algumas dessas alternativas devem ser mencionadas.

Testemunhos agradecidos. Considerando o que representam a gentileza e a gratidão humanas, a forma mais segura de se garantir uma avaliação favorável é obtida através da utilização de testemunhos voluntários daqueles que receberam o tratamento. Se os testemunhos surgidos espontaneamente foram escassos, estes devem ser solicitados entre os recipientes com os quais o programa ainda mantém contato. O otimismo que esses testemunhos inspiram é análogo à impressão que um professor tem do seu sucesso no ensino quando ouve comentários apenas dos alunos que vêm procurá-lo e conversar com ele após a aula. Em muitos programas, como na psicoterapia, o recipiente, assim como a unidade administrativa, gasta muito tempo e esforço com o programa. Nesse caso o comunicado de uma melhora, além de reduzir o sentimento de frustração, é uma gentileza para com o terapeuta. Os testemunhos agradecidos podem vir na linguagem das cartas e de conversas, ou enquadrados nas respostas a um "teste" de múltipla escolha, nos quais um tema freqüente é "estou doente", "estou bem", "estou feliz" e "estou triste". É provável que o testemunho seja tanto mais favorável: a) quanto mais claro for para o recipiente o caráter de avaliação da resposta — é perfeitamente claro na maioria dos testes de personalidade, ajustamento, moral e atitude; b) quanto mais direta for a identificação do

nome do recipiente que responde à pergunta; c) quanto mais o recipiente dá sua resposta diretamente ao terapeuta ou agente da reforma; d) quanto mais o agente continue a ser influente na vida futura do recipiente; e) quanto mais as respostas lidam com sentimentos e avaliações em vez de lidarem com fatos verificáveis; e f) quanto mais os recipientes que participam na avaliação constituem um subgrupo pequeno dos recipientes, formado de voluntários ou de elementos escolhidos pelo agente. Se for bem planejado, o método do testemunho agradecido pode compreender testes prévios, além de testes posteriores, e envolver grupos de controle designados por aleatorização, pois geralmente não se usam pseudotratamentos e os recipientes sabem perfeitamente que eles foram beneficiados.

Confundir seleção e tratamento. Outra tática segura para se obter resultados favoráveis é confundir a seleção com o tratamento, de modo que na comparação levada ao conhecimento do público os que receberam o tratamento são também os mais capazes e bem colocados. A tão citada evidência do valor por dólar de uma educação em faculdade é desse tipo — todos os estudos cuidadosos mostram que a maior parte do efeito, e do efeito mais acentuado das melhores faculdades, pode ser explicado por um talento maior e por contatos familiares e não pelo que é aprendido ou mesmo pelo prestígio do título. As técnicas de emparelhamento e do parcelamento estatístico não fornecem, em geral, um controle eficaz das diferenças de seleção, pois introduzem efeitos de regressão que podem ser confundidos com os efeitos do tratamento.

Temos que distinguir dois tipos de situação. Em primeiro lugar, existem aqueles tratamentos que são ministrados aos mais promissores, como a educação em faculdade, que normalmente é dada aos que menos precisam dela. Para esses tratamentos, as circunstâncias concomitantes com os motivos da seleção e que se manifestam mais tarde agem no mesmo sentido do tratamento: as que têm mais probabilidade de sucesso com a educação, ou sem ela, têm também mais probabilidade de entrar numa faculdade para depois conseguir sucesso. Para essas situações, o administrador encurralado deveria usar a média geral de todos os que receberam o tratamento e compará-la com a média de todos os que não o receberam, embora neste caso quase que qualquer comparação que pudesse ocorrer a um administrador seria viciada em seu favor.

No outro extremo da escala de talento estão os tratamentos corretivos ministrados àqueles que mais precisam dele. Neste caso, as circunstâncias concomitantes com os motivos de seleção e que se manifestam mais tarde são um menor sucesso. No exemplo do Corpo de Treinamento para Empregos, uma comparação descuidada da taxa de desemprego posterior dos que receberam o treinamento com a dos que não o receberam é, em geral, viciada contra o efeito do treinamento. O administrador encurralado deve ter cuidado neste caso e procurar aquelas poucas comparações especiais que viciam a seleção a seu favor. Para programas de treinamento tais como a Operação "Head Start" e programas de aulas particulares, uma solução útil é comparar o sucesso posterior dos que completaram o programa de treinamento com o dos que foram convidados mas nunca apareceram e, também o dos que vieram algumas vezes e abandonaram o programa. Considerando como "treinados" somente os que terminam o programa e usando os outros como controles, está-se fazendo uma seleção com base no grau de consciência individual, numa base familiar estável e amparadora, no gosto pela atividade de treinamento, na aptidão, na resolução de vencer na vida — fatores todos que prometem sucesso futuro mesmo que o programa corretivo não tenha valor algum. Para aplicar eficazmente esta tática no Corpo de Treinamento para Empregos seria necessário, talvez, eliminar do pretenso grupo de controle todos os que abandonaram o programa de treinamento porque encontraram um emprego — mas isto pareceria ser um procedimento razoável e não macularia o recebimento de um jubiloso relatório de andamento.

Essas são só mais duas amostras de modos de análise infalíveis para o administrador que não pode fazer face a uma avaliação honesta da reforma social que ele dirige. Esses exemplos nos fazem lembrar novamente que temos que ajudar a criar um clima político que exija testes da realidade mais rigorosos e menos enganosos. Devemos criar posturas políticas que permitam experimentos verdadeiros ou bons quase-experimentos. Das várias sugestões visando a esse objetivo dadas neste artigo, a mais importante é provavelmente o tema inicial: os administradores e os partidos políticos devem preconizar a importância do problema e não a importância de uma solução. Eles devem preconizar seqüências experimentais de reformas em vez de uma panacéia infalível, propondo uma Reforma *A* e tendo uma Alternativa *B* disponível para ser experimentada em seguida, no caso em que uma avaliação honesta de *A* mostrasse que ela fora inútil ou prejudicial.

8. REPETIÇÃO MÚLTIPLA DA AVALIAÇÃO

Número excessivo de cientistas sociais espera que um único experimento resolva uma questão definitivamente. Isto pode ser uma generalização errônea da história dos grandes experimentos cruciais da física e da química. Na realidade, os experimentos significativos das ciências físicas são repetidos milhares de vezes, não somente em esforços deliberados de repetição, mas também como eventualidades inevitáveis na experimentação sucessiva e na utilização dos muitos dispositivos de medição (como o galvanômetro), que na sua operação incorporam os princípios dos experimentos clássicos. Devido ao fato de que nós, cientistas sociais, possuimos menos poder para conseguir "isolamento experimental", já que temos boas razões para esperar que os efeitos do tratamento interajam significativamente com uma grande variedade de fatores sociais, muitos dos quais não levamos ainda em consideração, nossa necessidade de experimentos de repetição é muito maior do que a do cientista físico.

As implicações são claras. Devemos ser obstinados no teste da realidade não só na avaliação do programa-piloto e escolha da reforma a ser implementada na forma de lei. Devemos também, desde o momento em que tenha sido decidido que a reforma será adotada como padrão em todas as unidades administrativas, avaliar

experimentalmente os efeitos da reforma em cada uma das suas implementações (Campbell, 1967).

9. CONCLUSÕES

Os *administradores encurralados* já estão antecipadamente tão comprometidos com a eficácia da reforma que eles não se podem permitir uma avaliação honesta dos resultados. Recomendam-se para eles análises tendenciosas a seu favor, inclusive tirando proveito da regressão, dos testemunhos agradecidos e da confusão entre seleção e tratamento. Já os *administradores experimentais* terão justificado a reforma com base na importância do problema e não na certeza da solução, e seu compromisso é tentar outras possíveis soluções se a primeira falhar. Eles não estão, portanto, ameaçados por uma análise perspicaz da reforma. Por meio de decisões administrativas adequadas, podem estabelecer a base para úteis análises experimentais ou quase-experimentais. Com a ideologia de distribuir recursos escassos por sorteio, com o uso da inovação gradativa e com projetos-pilotos, poderão conseguir experimentos verdadeiros com grupos de controle de designação aleatória. Se a reforma precisar ser introduzida simultaneamente em todas as unidades administrativas, podemos usar o esquema da série temporal interrompida. Se houver unidades semelhantes sob administração independente, um esquema de série de controle reforça a análise. Se um benefício escasso precisa ser distribuído aos que mais necessitam dele ou aos que mais o merecem, a quantificação dessa necessidade ou desse mérito torna possível a análise da descontinuidade na regressão. ■

BIBLIOGRAFIA

Aubert, V. Chance in social affairs. *Inquiry*, n. 2, p. 1-24, 1959.

Bauer, R. M. *Social indicators*. Cambridge, Mass. M.I.T. Press, 1966.

Blayney, J. R. & Hill, I.N. Fluorine and dental caries. *The Journal of the American Dental Association*, (número especial),v. 74, p. 233-302, 1967.

Box, G.E.P. & Tiao, G.C. A change in level of a non-stationary time series. *Biometrica*, v. 52, p. 181-92, 1965.

Campbell, D.T. Factors relevant in the validity of experiments in social settings. *Boletim Psicológico*, v. 54, p. 297-312, 1957.

Campbell, D.T. From description to experimentation: interpreting trends as quasi-experiments. In: Harris, C.W., ed. *Problems in measuring change*. Madison. University of Wisconsin Press, 1963.

Campbell, D.T. Administrative experimentation, institutional records and nonreactive measures. In: Stanley, J.C., ed. *Improving experimental design and statistical analysis*. Chicago. Rand McNally, 1967.

Campbell, D.T. Quasi-experimental design. In: Sills, D.L., ed. *International encyclopedia of the social sciences*. New York. Macmillan and Free Press, 1968, v. 5, p. 259-63.

Campbell, D.T. & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, v. 56, p. 81-105, 1959.

Campbell, D.T. & Ross, H.L. The Connecticut crackdown on spending: time-series data in quasi-experimental analysis. *Law and Society Review*, v. 3, n. 1, p. 33-53, 1968.

Campbell, D.T. &. Stanley, J.C. Experimental and quasi-experimental designs for research on teaching. In: Gage, N.L., ed. *Handbook of research on teaching*. Chicago. Rand McNally, 1963. (Reeditado como *Experimental and quasi-experimental design for research*. Chicago. Rand McNally, 1966.)

Chapin, F.S. *Experimental design in sociological research*. New York. Harper, 1947.

Etizioni, A. "Shortcuts" to social change? *The Public Interest*, v. 12, p. 40-51, 1968.

Etzioni, A. & Lehman, E.W. Some dangers in "valid" social measurement. *Annals of the American Academy of Political and Social Science*, v. 373, p. 1-15, 1967.

Galtung, J. *Theory and methods of social research*. Oslo. Universitetsforloget; London. Allen and Unwin; New York. Columbia University Press, 1967.

Glass, G.V. Analysis of data on the Connecticut speeding crackdown as a time-series quasi-experiment. *Law and Society Review*, v. 3, n. 1, p. 55-76, 1968.

Glass, G.V.; Tiao, G.C. & Maguire, T.O. Analysis of data on the 1900 revision of the German divorce laws as a quasi-experiment. *Law and Society Review*, no prelo.

Greenwood, E. *Experimental sociology: a study in method*. New York. King's Crown Press, 1945.

Gross, B.M. *The state of the nation: social system accounting*. London. Tavistock Publications, 1966. (Também em R.M. Bauer. *Social indicators*. Cambridge. Mass. M.I.T. Press, 1966.)

Gross, B.M., ed. Social goals and indicators. *Annals of*

the American Academy of Political and Social Science, v. 371, Parte 1, May, p. i-iii e 1-177; Parte 2, Sept. p. i-iii e 1-218, 1967.

Guttman, L. An approach for quantifying paired comparisons and rank order. *Annals of Mathematical Statistics*, v. 17, p. 144-63, 1946.

Hyman, H.H. & Wright, C.R. Evaluating social action programs. In: Lazarsfeld, W.H. Sewell & Wilensky, H.L., ed. *The uses of sociology*. New York. Basic Books, 1967.

Kamisar, Y. The tactics of police-persecution oriented critics of the courts. *Cornell Law Quarterly*, v. 49, p. 458-71, 1964.

Kaysen, C. Data banks and dossiers. *The Public Interest*, v. 7, p. 52-60, 1967.

Manniche, E. & Hayes, D.P. Respondent anonymity and data matching. *Public Opinion Quarterly*, v. 21, n. 3, p. 384-8, 1957.

Office of the Secretary of Defense, Assistant Secretary of Defense (Manpower). Guidance paper: Project One Hundred Thousand. Washington, D.C., March 31, 1967 (multilith).

Polanyi, M. A society of explorers. In: *The tacit dimension*. New York. Doubleday, 1966. Cap. 3.

Polanyi, M. The growth of science in society. *Minerva*, v. 5, p. 533-45, 1967.

Popper, K.R. *Conjectures and refutations*. London. Routledge and Kegan Paul; New York. Basic Books, 1963.

Rheinstein, M. Divorce and the law in Germany: a review. *American Journal of Sociology*, v. 65, p. 489-98, 1959.

Rose, A.M. Needed research on the mediation of labor disputes. *Personnel Psychology*, v. 5, p. 187-200, 1952.

Ross, H.L. & Campbell, D.T. The Connecticut speed crackdown: a study of the effects of legal change. In: Ross, H.L. ed. *Perspectives on the social order: readings in sociology*. New York. MacGraw-Hill, 1968.

Sawyer, J. & Schechter, H. Computers, Privacy and the National Data Center: the responsibility of social scientists. *American Psychologist*, v. 23, p. 810-18, 1968.

Schanck, R.L. & Goodman, C. Reactions to propaganda on both sides of a controversial issue. *Public Opinion Quarterly*, v. 3, p. 107-12, 1939.

Schwartz, R.D. Field experimentation in sociological research. *Journal of Legal Education*, v. 13, p. 401-10, 1961.

Schwartz, R.D. & Orleans, S. On Legal sanctions. *University of Chicago Law Review*, v. 34, p. 247-300, 1967.

Schwartz, R.D. & Skolnick, J.H. Televised comunication and income tax compliance. In: Arons, L. & May, M., ed. *Television and human behavior*. New York. Appleton-Century-Crofts, 1963.

Selvin, H. A critique of tests of significance in survey research. *American Sociological Review*, v. 22, p. 519-27, 1957.

Simon, J.L. The price elasticity of liquor in the U.S. and a simple method of determination. *Econometrica*, v. 34, p. 193-205, 1966.

Solomon, R.W. An extension of control group design. *Psychological Bulletin*, v. 46, p. 137-50, 1949.

Stieber, J.W. *Ten years of the Minnesota Labor Relations Act*. Minneapolis. Industrial Relations Center, University of Minnesota, 1949.

Stouffer, S.A. The point system for redeployment and discharge. In: Stouffer, S. A. et alii. *The American soldier*. Vol. 2, *Combat and its aftermath*. Princeton. Princeton University Press, 1949.

Suchman, E.A. *Evaluative research: principles and practice in public service and social action programs*. New York. Russell Sage, 1967.

Sween, J. & Campbell, D.T. A study of the effect of proximally auto-correlated error on tests of significance for the interrupted time-series quasi-experimental designs. Recebida do autor, 1965 (multilith).

Thistlethwaite, D.L. & Campbell, D.T. Regression-discontinuity analysis: an alternative to the ex post-facto experiment. *Journal of Educational Psychology*, v. 51, p. 309-17, 1960.

Walker, H.M. &. Lev, J. *Statistical inference*. New York. Holt, 1953.

Webb, E.J.: Campbell, D.T.; Schwartz, R.D. &. Sechrest, L.B. *Unobstrusive measures: nonreactive research in the social sciences*. Chicago. Rand McNally, 1966.

Wolf, E.; Lüke, G. & Max, H. *Scheidung und Scheidungsrecht: Grundifrägen der Ehescheidung in Deutschland*. Tubigen. J.C.B. Mohr, 1959.

1 Esta lista foi ampliada em relação às versões anteriores com a adição de *Instabilidade* (ver também Campbell, 1968; Campbell e Ross, 1968). Esta adição foi feita como reação à discussão sociológica do uso dos testes de significância na pesquisa não-experimental e quase-experimental (Selvin, 1957; a crítica desse trabalho feita por Galtung, 1967, p. 358-89). Por um lado, uno-me aos que criticam o prestígio exagerado das "diferenças estatisticamente significativas" no estabelecimento de certeza de validade. Na melhor das hipóteses, os testes estatísticos só são relevantes para 1 dentre 15 ameaças à vali-

dade. Por outro, concordo com os que defendem seu uso em situações onde não foi usada a aleatorização. Mesmo nesses casos, faz sentido dizer-se ou negar-se: "Esta diferença é trivial. É da ordem que teria ocorrido com freqüência *se* essas medidas *tivessem* sido designadas por pura chance." Os testes de significância que utilizam uma redesignação aleatória das medidas realmente obtidas são especialmente úteis para se transmitir este argumento.

2 Esta lista também foi ampliada em relação às versões anteriores deste trabalho para tornar mais evidentes as ameaças 5 e 6, as quais são especialmente relevantes na experimentação social. A discussão nas versões anteriores (Campbell, 1957, p. 309-310; Campbell e Stanley, 1963, p. 203-4) tinha abrangido essas questões mas não haviam sido incluídas na lista de verificação.

3 Não há dúvida de que tanto o público como a imprensa participaram do susto do governador com o número de mortes em 1955. Essa reação discriminatória poderia ser encarada como um sistema de realimentação negativa no qual o efeito amortecedor é proporcional ao aumento em relação à tendência prévia. Na medida em que tal susto causa uma redução nas mortes causadas pelo trânsito, ele acrescenta uma componente negativa à autocorrelação, aumentando o efeito de regressão. Esta componente deveria provavelmente ser encarada como uma causa rival ou um tratamento rival em vez de como um efeito ilusório. (O efeito de regressão é menor quanto maior for a autocorrelação positiva e estará presente na medida em que essa correlação é menor do que a unidade positiva. Uma correlação negativa numa série temporal representaria uma regressão além da média, numa forma não exatamente análoga à correlação negativa entre pessoas. Para autocorrelação com retardamento 1, uma alta correlação negativa seria representada por uma série que oscilasse com máxima amplitude de um extremo a outro.)

4 A inconsistência de Wilson na utilização dos registros e o problema político de registros relevantes estão competentemente documentados em Kamisar (1964). Etzioni (1968) relata que em 1965, em Nova York, foi proclamada uma onda de crimes que se revelou depois ser devida a uma melhora não divulgada no sistema de registro.

5 Sween, J. e Campbell, D. T. *Computer programs for simulating and analyzing sharp and fuzzy regression-discontinuity experiments.* Em preparação.

6 Embora disponhamos de pelo menos um teste de significância exeqüível, pode ser bem difícil conseguir um teste que preserve a imagem de se extrapolar para um hipotético teste de desempate com aleatorização. Inicialmente, seguindo a orientação de Walker e Lev (1953, p. 400; Sween e Campbell, 1965, p. 7), testamos a significância da diferença das duas linhas de regressão no ponto de separação, uma ajustada às observações abaixo do ponto de separação e a outra ajustada às observações acima dele. Na simulação por computador de casos de efeito nulo, foram encontrados repetidamente pseudo-efeitos "significativos". Acontece que esta é uma daquelas situações em que a solução pelo método dos mínimos quadrados é viciada. Uma forma de compreender a natureza desse vício é talvez considerar o que aconteceria se tanto a reta de regressão do teste prévio sobre o teste posterior como a do teste posterior sobre o teste prévio fossem traçadas para toda a distribuição. Essas duas retas de regressão cruzariam no centro da distribuição (isto é, no ponto de separação, em exemplos simétricos como os das figuras 14 e 15) e se afastariam nas extremidades. Quando, em vez disso, as duas retas de regressão são ajustadas para cada metade da distribuição, elas cruzarão no centro de cada metade e se afastarão nas imediações do ponto de separação. Num exemplo como o da figura 14, a regressão do teste posterior sobre o teste prévio será a mais baixa no ponto de separação para a metade não tratada e a mais alta para a metade tratada. Este pseudo-efeito não aparece quando se traçam os pontos representando as médias de cada coluna, o que pode ser verificado visualmente, e as figuras 14, 15, 16 e 17 deveriam ter sido desenhadas com as médias de cada coluna representadas em vez das retas ajustadas. O tamanho desse vício é uma função da correlação entre o teste prévio e o teste posterior e se esta puder ser adequadamente estimada, poder-se-ia calcular uma estimativa corrigida da diferença no ponto de separação. No entanto, não se pode usar a distribuição inteira para se estimar essa correlação, pois um efeito real irá causar parte da correlação. Poder-se-ia basear uma estimativa nas correlações calculadas em separado para as partes acima e abaixo do ponto de separação, corrigindo-a pelo fato de abrangerem uma faixa restrita. Poder-se-ia também encontrar procedimentos de estimação de máxima verossimilhança.

No momento, a melhor sugestão parece ser a que foi fornecida por Robert P. Abelson. A reta de regressão do teste posterior sobre o teste prévio é ajustada para um grupo de dados que se estendem para cima e para baixo do ponto de separação em porções iguais. As médias das colunas são expressas como desvios daquela regressão. Um teste *t* é então usado para se comparar as colunas junto ao ponto de separação, acima e abaixo dele. Para aumentar a base estatística, pode-se explorar uma classificação em colunas mais largas. Este teste infelizmente perde a analogia com o experimento verdadeiro de desempate, analogia da qual o presente autor lançou mão para um esclarecimento conceitual.

7 Há alguns indícios estatísticos sutis que poderiam distinguir estes dois casos se tivéssemos observações suficientes. Deveria haver um aumento da variância dos valores combinados das colunas nas colunas mistas no caso de um efeito real. Se os dados fossem tratados arbitrariamente como se tivesse havido um ponto de separação nítido no meio da região em que as observações se misturam, então não deveria haver descontinuidade no caso de efeito nulo e sim alguma descontinuidade no caso de efeito real, embora neste segundo caso a descontinuidade fosse subestimada, já que existem casos não tratados acima do ponto de separação e casos tratados abaixo desse ponto, diminuindo o efeito visível. A intensidade desta diminuição deveria ser estimável e corrigível, talvez através de procedimentos iterativos. Mas estas são esperanças para o futuro.