Manuel Souto Otero _Universidade de Oxford

Administración&Cidadanía. _1/2006

Avaliación de políticas: unha revisión crítica das definicións, deseños e métodos

***A bibliografía sobre** a avaliación de políticas é extremadamente voluminosa e desordenada. Este artigo pretende achegar un resumo crítico das definicións, plans e métodos utilizados na avaliación de políticas para aqueles lectores que estean interesados neste tema pero que se sintan desbordados e, loxicamente, confusos debido á variedade de enfoques existente. O artigo proporciona tamén unha lista dos elementos que caracterizan unha boa avaliación de políticas, que quere ser de utilidade para os avaliadores e os creadores de políticas e para outras persoas implicadas no uso e exame das avaliacións de políticas.&*

& Palabras clave: avaliación de políticas públicas, elaboración de políticas públicas, impacto de políticas públicas, métodos de investigación

1. QUE É A AVALIACIÓN DE POLÍTICAS?

A análise das políticas públicas pretende determinar cal das varias políticas públicas ou gobernamentais posibles acadará en maior medida unhas metas determinadas, en vista das relacións entre as políticas e as metas (Nagel, 1990). Nesta tarefa iterativa, a avaliación de políticas desempeña un papel determinante para axudarlles aos xestores a deseñar ou mellorar o deseño de políticas, programas e iniciativas, e para fornecer, onde cumpra, de exames periódicos da eficacia de políticas ou programas, dos seus impactos e dos xeitos alternativos de alcanzar os resultados esperados.

Xa que logo a avaliación, do mesmo xeito ca as auditorías internas, a xestión de riscos e outras ferramentas, axuda aos xestores a traballar eficazmente en contornos políticos sumamente complexos. Pero a avaliación tamén pode axudarlles aos xestores no control do rendemento e informar sobre el, e axudarlles aos responsables da toma de decisións a examinar os resultados dos programas ou as políticas. Isto é o que distingue a avaliación da auditoría interna, unha función que proporciona garantías sobre a estratexia de xestión de riscos dun departamento ou axencia, un marco para o control de xestión e información, financeira e non financeira, utilizada para a toma de decisións e a elaboración de informes (Treasury Board of Canada Secretariat, 2001).

A avaliación de políticas pode definirse como un conxunto de métodos de investigación que pretende "investigar sistematicamente a eficacia das intervencións sociais (...) de xeito que melloren as condicións sociais" (Rossi *et al.,* 1999: 20). A importancia da noción de conxunto de métodos de investigación é fundamental nesta definición: a avaliación de políticas utiliza unha ampla gama de deseños e métodos sen privilexiar ningún; pola contra, recoñece o potencial complementario de diversos métodos de investigación. Os métodos utilizados na avaliación e análise das políticas derivan xeralmente dos temas importantes que se tratan, máis que por preferencias *a priori*[1] (Greene *et al.,* 2001).

Este artigo proporciona unha revisión dos diferentes tipos e métodos de avaliación. Agora ben, o lector debería recordar que, aínda que ampla, non quere ser exhaustiva senón que pretende ofrecer un resumo e análise dos principais enfoques de avaliación. Ademais, algúns dos tipos e métodos de avaliación presentados aquí son complementarios e outros alternativos; uns poden aplicarse de xeito xeral mentres outros teñen aplicacións bastante específicas. Este artigo estrutúrase como segue: a sección segunda presenta unha gama de diversos tipos de avaliacións e plans de avaliación, con particular énfase nas avaliacións conclusivas, a forma de avaliación máis estendida, mentres que a sección terceira céntrase nos métodos de recolección de datos que poden utilizarse para levar a cabo estes diferentes tipos de avaliación; pola súa banda, a cuarta sección conclúe ofrecendo algúns criterios que cómpre satisfacer nunha boa avaliación de políticas.

2. TIPOS DE AVALIACIÓN

2.1. Introdución

Nalgúns países europeos hai pouca tradición de avaliación de políticas: estas son deseñadas, postas en práctica e cambiadas sen estudar que resultado poderían dar e cal deron realmente. Noutros estados, a tradición da avaliación está moito máis desenvolvida, ata o punto

de que as políticas son avaliadas case por defecto, independentemente das características específicas das diversas políticas, programas ou proxectos. Ningunha destas situacións é a ideal.

A primeira pregunta que hai que formular no intre de decidir a avaliación dunha política, un programa ou un proxecto é a seguinte: pode sequera avaliarse esta intervención? Algunhas políticas e algúns programas son tan complexos e difusos que hai poucas posibilidades de que satisfagan os tres requisitos principais que posibilitan a avaliación: que as intervencións e a poboación a que están destinadas estean claras e sexan identificables; que os resultados sexan claros, específicos e conmensurables; e que se poida poñer en práctica un plan de avaliación apropiado (Patton, 2002).

Se se satisfán estes criterios, a avaliación de políticas tería que producir beneficios significativos. Neste senso, os gobernos, as organizacións internacionais e demais partes implicadas fan uso dunha gama de diversos tipos de avaliación, dependendo dos seus obxectivos, do seu enfoque analítico favorito e do estadio de desenvolvemento concreto da política sobre a que se require información.

En adiante preséntanse varios tipos de avaliación, que inclúen métodos de valoración rápida, análise lóxica do contexto, indicadores de rendemento, avaliacións antes, durante e despois da posta en práctica, avaliacións baseadas nas metas e independentes delas, avaliacións baseadas na teoría, avaliacións cuantitativas e cualitativas, avaliacións formativas e conclusivas, informes de seguimento do gasto público, a análise custo-beneficio, custo-efectividade e custo-utilidade. Como xa se fixo notar, algúns destes tipos de avaliación son complementarios e outros exclusivos. Daquela, a división entre tipos que aquí se presenta non pretende ser unha clasificación de tipos de avaliación: por exemplo, non hai un *fundamentum divisionis* compartido (Marradi, 1990) por todos os tipos de avaliación aquí presentados.

2.2. Métodos de valoración rápida, marco lóxico e indicadores de rendemento

Os modelos presentados nesta subsección non son métodos exhaustivos de avaliación, mais teñen en común coas avaliacións a intención de contribuír aos procesos de toma de decisións.

Os primeiros métodos, os métodos de valoración rápida, son formas rápidas e baratas de reunir os puntos de vista e os comentarios dos beneficiarios dunha política, así como doutras partes interesadas relevantes, para dar resposta á necesidade de información dos responsables da decisión. Proporcionan información, normalmente cualitativa, recompilada rapidamente para a toma de decisións dos xestores, especialmente no nivel dos proxectos e programas, aínda que as súas conclusións non adoitan poder xeneralizarse e son menos válidas, fidedignas e cribles ca as que se obteñen por medio das avaliacións. Os métodos de valoración rápida poden incluír entrevistas con informadores clave (unha serie de preguntas amplas formuladas a individuos seleccionados polo seu coñecemento e experiencia respecto dun tema), discusións de grupo (un debate entre, polo xeral, de oito a doce participantes con experiencias similares, como por exemplo os beneficiarios ou o persoal do programa), entrevistas en grupos de comunidade (unha serie de preguntas e un debate que teñen lugar nun encontro aberto a todos os membros da comunidade), observación directa (sistema baseado no uso dun impreso de observación detallado para rexistrar o que se ve e o que se oe no escenario dun programa) ou mini-informes (que usan un cuestionario estruturado cun número limitado de preguntas concretas que se lle presenta a un número pequeno de persoas escollidas ao chou ou seguindo algún criterio).

O marco lóxico vai un paso máis alá dos métodos de valoración rápida. Pretende axudar a aclarar os obxectivos de calquera proxecto, programa ou política e identificar as relacións causa-efecto que cabe esperar –a 'lóxica do programa'– na seguinte concatenación de resultados: recursos, procesos, resultados, consecuencias e impacto (véxase a sección 2.6 para unha definición destes elementos). A análise lóxica do contexto é, entón, un xeito de implicar os interesados na concreción dos obxectivos e o deseño de actividades, e conduce á identificación de indicadores de rendemento en cada fase desta concatenación, así como á identificación dos riscos que poderían impedir o cumprimento destes obxectivos.

Pola súa parte, os indicadores de rendemento son medicións de recursos, proceso, resultados, consecuencias e impacto dos proxectos, programas ou estratexias de desenvolvemento. Cando os apoia unha recompilación sólida de datos e un informe sobre eles, os indicadores permítenlles aos xestores controlar os progresos, demostrar os resultados e adoptar accións correctivas para mellorar a prestación de servizos. O uso de indicadores de rendemento implica o establecemento de metas e a valoración dos progresos que se fan para as acadar. Teñen a vantaxe de que, por medio do seu uso, se foron seleccionados con rigor, é posible identificar problemas mediante un sistema de alarmas temperás que permite que se dean pasos para corrixilos. Tamén poden indicar se unha determinada intervención necesita unha avaliación ou un informe exhaustivos desde o principio.

2.3. Avaliacións ex-ante, intermedias e ex-post

Non cómpre dicir que non todas as avaliacións se efectúan ao final da posta en práctica dunha política, programa ou proxecto: o proceso de avaliación debería contribuír a melloras na política, así como no deseño de programas e na súa posta en práctica. Rossi, por exemplo, suxire que se poñan en práctica métodos deseñados a medida da fase en que se atopa o programa, é dicir, que se adecúen 'as avaliacións ao programa' (Rossi *et al.,* 1999). Xa que logo, ademais das avaliacións ex-post, existen as ex-ante e as intermedias.

As avaliacións ex-ante teñen lugar antes do comezo dunha política, programa ou proxecto e contribúen a establecer claramente a lóxica que se seguiría para intentar resolver un problema e os métodos para facelo, así como os efectos positivos e negativos que se espera que a intervención provoque. Normalmente téñense en conta a relación e a coherencia entre obxectivos globais, obxectivos específicos e a complementariedade das medidas que se incluirán no programa; tamén se dedica unha parte importante do traballo a desenvolver un sistema apropiado de indicadores que se utilizarán en avaliacións subsecuentes (Comisión Europea, 1999). Durante as avaliacións *ex-ante* pode terse en conta toda unha gama de posibles intervencións; por conseguinte, as avaliacións previas son unha boa oportunidade para mellorar a planificación de políticas e poden sustentar con éxito discusións respecto diso desde varios puntos de vista.

As avaliacións intermedias realízanse durante o desenvolvemento dunha determinada intervención, e normalmente céntranse no seu transcurso ata un punto dado, salientando particularmente os estadios de planificación e posta en práctica da iniciativa en cuestión e prestándolles menos atención aos resultados ou consecuencias, que poden ser difíciles de identificar nunha fase temperá da posta en práctica. Examinan o grao de efectividade conseguido baseándose nos indicadores recompilados durante o proceso de seguimento da iniciativa e valoran a súa calidade e relevancia. As avaliacións ex-ante e intermedias non adoitan ser un fin en si mesmas senón unha forma de mellorar a calidade e a relevancia dun progra-

ma: proporcionan unha oportunidade para identificar as reorientacións do propio programa que poden facer falla para garantir que se consigan os obxectivos orixinais (Comisión Europea, 2000).

A pesar de todo, as avaliacións ex-post son o tipo máis frecuente: examinan se se alcanzaron os resultados esperados dunha intervención e subministran información necesaria para a planificación ou posta en práctica de programas novos ou revisados. As avaliacións ex-post adoitan usar datos definitivos de seguimento e comparar os obxectivos esperados cos realmente conseguidos, incluíndo o impacto (Comisión Europea, 1999). Aínda así, non todas as avaliacións posteriores se centran exclusivamente na análise dos resultados acadados en relación coas metas fixadas para a intervención. Así pois, podemos distinguir entre avaliacións baseadas nas metas, avaliacións independentes das metas e avaliacións baseadas na teoría.

2.4. Avaliación de metas e avaliacións libres

Unha das preguntas máis frecuentes na avaliación de políticas é se se cumpriron ou non as metas dunha política, dun programa ou dun proxecto: o exame que responde a esta pregunta coñécese como avaliación de metas (Patton, 2002). Na bibliografía estadounidense sobre as avaliacións coñécese tamén como 'seguimento lexislativo', porque analiza se se alcanzaron os resultados esperados dunha política gobernamental.

Con todo, os creadores de políticas e os avaliadores están interesados acotío tamén nos resultados ou efectos imprevistos, positivos ou negativos, dunha política, programa ou proxecto, aínda sen saberen necesariamente cales eran as metas prefixadas. Este tipo de avaliación de políticas é a miúdo importante para establecer a relación entre custo e beneficio ou entre custo e utilidade dunha política, un programa ou unha intervención. Scriven (1972), por exemplo, propugna a 'avaliación libre', porque o avaliador asume a responsabilidade de decidir que resultados do programa examinar e rexeita tomar como punto de partida os obxectivos deste. Mantén que, obrando así, o avaliador está máis capacitado para identificar os verdadeiros logros (e fracasos) do programa.

2.5. Avaliacións baseadas na teoría

As avaliacións de metas, e con menor frecuencia as libres, poden ou non estar rexidas pola teoría. Os modelos de avaliación baseados na teoría –que inclúen o modelo de teorías do cambio, así como a avaliación da teoría do programa (Weiss, 1997; Rodgers *et al.,* 2000) e algúns aspectos da avaliación realista (Pawson e Tilley, 1997)– non se distinguen polos resultados que tratan de explicar senón polo seu interese en descifrar a secuencia lóxica ou teórica pola que se espera que unha intervención dea os resultados desexados. Localizando os factores determinantes ou causais que se consideran importantes para o éxito, e como poden interactuar, pódese decidir que pasos deberían seguirse con maior atención a medida que se desenvolve o programa (Rossi *et al.,* 1999).

Huey-Tsyh Chen, un dos autores máis influentes no desenvolvemento do concepto e da práctica da avaliación baseada na teoría, argumentou que unha desafortunada consecuencia das avaliacións non baseadas na teoría (como moitas das que utilizan probas efectuadas ao azar; véxase máis adiante) é que os resultados da avaliación achegan visións frecuentemente limitadas e por veces deformadas dos programas (Chen e Rossi, 1983). As teorías que desexa construír non son globais nin ambiciosas senón "modelos verosímiles e defendibles de como se pode esperar que funcionen os programas" (Chen e Rossi, 1983), nos cales basear o exercicio de avaliación en calquera estadio dunha intervención dada.

2.6. Avaliacións formativas e conclusivas

2.6.1. Introdución

Tamén se coñece a avaliación formativa como avaliación do proceso e a avaliación conclusiva, como avaliación do impacto. As avaliacións intermedias son xeralmente formativas, aínda que tamén poden ter elementos aditivos; as avaliacións ex-post –véxase máis arriba– adoitan ser conclusivas, aínda que tamén poden examinar os procesos subxacentes a unha intervención.

A avaliación formativa pregunta como, por que e en que circunstancias funciona ou non unha política, un programa ou un proxecto: estas preguntas son importantes á hora de determinar a posta en práctica eficaz das políticas, os programas ou os proxectos. Este tipo de avaliación adoita buscar información nos factores, mecanismos e procesos contextuais que están detrás do éxito ou do fracaso dunha política. Isto implica decote formular preguntas como para quen funcionou ou non unha política e por que.

As avaliacións conclusivas formulan preguntas como que impacto, de existir algún, ten unha política, un programa ou un proxecto sobre diversos grupos de xente: ten por obxectivo contrastar os efectos dunha política –duros ou brandos[2] (Lloyd e O'Sullivan, 2004)– co que se esperaba dela no estadio do seu deseño, con outra intervención ou coa ausencia de intervención (contrafactual). A avaliación do impacto é, daquela, a identificación sistemática dos efectos –positivos ou negativos, desexados ou non– causados por unha determinada intervención. A avaliación do impacto axuda a entender mellor ata que punto as actividades alcanzan os grupos aos que están destinadas e a magnitude dos seus efectos.

A distinción entre avaliacións conclusivas e formativas, no entanto, non é tan clara como se podería deducir destas definicións. Por exemplo, os partidarios do modelo de teorías do cambio (Chen, 1990; Cornell e Kubisch, 1995; Funnel, 1997; Owen e Rodgers, 1999; Weiss, 1997) sosteñen que, para determinar se unha política funcionou ou non ou se foi efectiva, é imprescindible formular preguntas sobre como funcionou, para quen, por que e baixo que condicións. Con todo, na bibliografía sobre avaliación de políticas adoita diferenciarse entre avaliar se unha política foi efectiva (avaliación conclusiva) e avaliar por que o foi (avaliación formativa). A seguir centrámonos nas avaliacións conclusivas ou de impactos, o tipo máis común de avaliación e no que maior diversidade metodolóxica pode atoparse.

As análises dos procesos e dos impactos tenden a seguir o seguinte modelo conceptual de intervencións mediante políticas: as administracións, as axencias ou os operadores poñen en marcha medidas utilizando diversos medios ou recursos, financeiros, humanos, técnicos ou organizativos. O investimento dá lugar a unha serie de resultados físicos –por exemplo, quilómetros de estradas construídos, número de centros de formación creados, etc.– que demostran os progresos xerados pola posta en práctica da medida. Os resultados son os efectos (inmediatos) nos beneficiarios directos das accións financiadas, por exemplo a redución da duración dos desprazamentos, os custos de transporte ou a cantidade de persoal formado. Estes resultados poden expresarse segundo os seus impactos á hora de conseguir os obxectivos globais ou específicos do programa, e forman a base principal para valorar o éxito ou o fracaso do servizo en cuestión. Os impactos específicos poden incluír, por exemplo, o incremento do tráfico de mercadorías ou unha formación máis acorde coas demandas do mercado laboral. Os impactos globais están relacionados coa meta xeral das axudas, como a

creación de emprego. Evidentemente, a medición deste tipo de impacto é complexa, e con frecuencia é difícil establecer relacións causais claras (Comisión Europea, 1999).

Os principais métodos para valorar o impacto pertencen a un destes grupos: métodos experimentais, que son esencialmente probas efectuadas ao azar, e métodos *quasi-experimentais*. Ambos intentan satisfacer a cuestión principal: estimar a adicionalidade das intervencións a partir dun cálculo do que sucedería de non existir o programa (contrafactual).

2.6.2. Métodos experimentais

A dificultade da estimación do contrafactual é evidente: nun momento dado obsérvase un individuo, afectado polo programa ou non. Na maioría dos casos, comparar o mesmo individuo a través do tempo non nos proporcionará unha estimación fiable do impacto que o programa poida ter sobre el, dado que desde que se introduciu o programa poden cambiar para el moitas máis cousas. Xa que logo, non podemos pretender obter un cálculo do impacto do programa en cada individuo; o máis que podemos esperar é poder obter o impacto medio do programa sobre un grupo de individuos comparándoo cun grupo similar que non estea exposto ao programa. Así pois, o obxectivo crítico da avaliación do impacto é establecer un *grupo de comparación*[3] crible, un grupo de individuos que *en ausencia do programa* experimentasen fenómenos similares aos daqueles que si estiveron expostos ao programa. Este grupo dános unha idea do que lle podería suceder ao grupo do programa se non estivese exposto a el e permítenos obter unha estimación do impacto medio no grupo en cuestión.

Considérase que os métodos de avaliación ao azar, que implican a recompilación de información sobre o grupo de intervención e sobre o grupo de comparación en dous ou máis momentos, ofrecen a análise máis rigorosa do impacto do proxecto e a contribución doutros factores. Na 'avaliación ao azar antes e despois do ensaio', ou 'avaliación do ensaio en circunstancias de azar controlado', os suxeitos –familias, escolas, comunidades, etc.– son atribuídos ao chou aos grupos de proxecto e control. O azar non garante que os dous grupos vaian ser idénticos, pero reduce a influencia dos factores extrínsecos, garantindo que as diferenzas entre os dous grupos estarán libres dun nesgo sistemático. Os ensaios efectuados en circunstancias de azar controlado abordan o problema de que outros factores posibles inflúan no resultado expoñendo o grupo de experimentación e o de control a exactamente as mesmas circunstancias, excepto a política, o proxecto ou o programa que se está investigando. Para unha discusión clásica sobre este tema, véxase Campbell e Stanley (1966).

Boruch (1997) argumenta que calquera programa, sexa social, educativo ou de benestar, debería ser estudado de xeito sistemático, empregando métodos experimentais en circunstancias de azar controlado para reunir probas válidas e fiables. Con todo, non está claro que se poidan avaliar todos os programas deste xeito: por exemplo, o exame dun tema como a independencia dun banco central tería que sustentarse sobre outros métodos de avaliación. Os programas dirixidos a individuos ou comunidades locais (como os servizos sanitarios, a reforma do goberno local, a educación e a sanidade), en cambio, son mellores candidatos a avaliacións aleatorias.

Aínda máis, os ensaios non permiten responder a todas as preguntas relacionadas coa avaliación: o seu problema principal é que dan unha idea aproximada da adicionalidade *neta*, pero non proporcionan maneira ningunha de distinguir a cantidade de xente para a que o programa mellora os efectos da cantidade de xente para a que os empeora. Este feito limita seriamente o punto ata o que se poden identificar os beneficios concretos do programa sobre

os individuos. Tamén hai que ter en conta que os ensaios aleatorios sobre individuos son de pouca utilidade se un dos obxectivos é producir un 'cambio de cultura' xeral que concirna a toda a poboación afectada: neses casos é case imposible evitar a contaminación do grupo de control. Para os programas deste tipo, a única opción factible para levar a cabo un ensaio ao azar pode ser a división aleatoria de áreas.

Aínda que o deseño de ensaios realizados en circunstancias de azar controlado é atractivo pola simplicidade, a execución pode ser complexa e require unha experiencia operativa e analítica considerable. Normalmente hai problemas éticos, suscitados polo feito de expoñer un grupo de xente (o experimental) a unha política potencialmente nociva ou, á inversa, polo feito de privar a outro grupo de xente (o de control) dunha política potencialmente beneficiosa. Non obstante, en ausencia de probas sólidas *a priori* de que unha política vaia ser nociva ou beneficiosa, adoita crerse que é eticamente aceptable efectuar un ensaio para dirimir este extremo, a condición de que se interrompa o ensaio en reuníndose probas válidas e fiables (Davies, 2004). Aínda así, moita xente opina que os métodos *quasi-experimentais* descritos no que resta deste artigo fornecen de resultados razoablemente fiables e evitan algúns dos problemas (relacionados coa ética e os recursos) que provocan os ensaios efectuados en circunstancias de azar controlado.

2.6.3. Métodos quasi-experimentais

Se se descarta o modelo de ensaio aleatorio ou se considera inapropiado, o reto para os avaliadores é elixir un método *quasi-experimental* alternativo que poida dar resultados razoablemente consistentes. Para isto hai unha gama de métodos. Neste artigo consideraremos:

- métodos de antes e despois
- métodos de intervalos cronolóxicos
- melloras de dobre diferenza
- métodos de comparación de dous grupos equivalentes
- modelado estatístico de datos xa existentes para a avaliación de programas voluntarios
- métodos de comparación de áreas equivalentes[4]
- análises económicas

2.6.3.1. Métodos de antes e despois

Nos métodos de 'antes e despois' identifícase a poboación á que afectará unha intervención antes e despois de que se introduza o programa. Selecciónase, de entre a poboación en cuestión, un 'grupo de programa' despois de que se introduza un programa e un 'grupo de comparación' con anterioridade á súa introdución (Greenberg e Morris, 2003). Despois recompílanse os resultados de ambos os grupos, e é a diferenza entre resultados o que proporciona o cálculo aproximado da adicionalidade. Neste método é da máxima importancia que no período previo ao programa se poida establecer quen son os candidatos elixibles. Este modelo utilizouse frecuentemente nas avaliacións, axustando, polo xeral, os resultados para poder controlar o efecto das características observables.

Os indicadores de antes e despois toman en consideración a selección de individuos baseada en características inobservables. A presuposición distintiva deste indicador é que a diferenza entre o contrafactual verdadeiro posterior ao programa e os resultados anteriores ao programa dá unha media de cero en todos os individuos participantes. Principalmente, o indicador de antes e despois presupón que as características inobservables son de dous tipos:

as particulares dun individuo estables no tempo (efectos individuais) e as particulares dun individuo pero non estables no tempo (efectos transitorios). Crese que a participación no programa depende do efecto fixo e non do transitorio. É unha presuposición moi forte que pode verse violada, por exemplo, polos cambios macroeconómicos entre os dous puntos de observación.

2.6.3.2. Métodos de intervalos cronolóxicos

O problema de como deslindar os cambios introducidos polo programa e os cambios históricos nos estudos de antes e despois pode afrontarse en certas ocasións ampliando o número de períodos anteriores e, de ser posible, o de períodos posteriores, de maneira que se obteña unha serie de intervalos cronolóxicos. Se nesta serie se dá unha ruptura no momento en que se introduce o novo programa ou pouco despois, interprétase que ese é o seu impacto. Os métodos baseados neste principio coñécense como 'métodos de intervalos cronolóxicos interrompidos'. Os intervalos cronolóxicos axudan a desbotar algunhas das posibles explicacións dun cambio; en particular, a condición de que non se introduzan programas relacionados a un tempo, un repentino cambio na serie resulta bastante concluínte. Isto é particularmente certo cando o cambio observado co programa é maior ca o observado entre calquera dos períodos anteriores. Se tamén se pode probar que o cambio que coincide coa introdución do programa perdura no tempo, as probas son máis sólidas: este é o motivo polo que interesa utilizar varios períodos posteriores.

Debido á necesidade de series de datos razoablemente longas nunha análise de intervalos cronolóxicos interrompidos, este método resulta máis indicado para a análise de datos administrativos, aínda que nalgúns casos poden utilizarse estudos de gran escala continuos ou case continuos.

No entanto, os métodos de intervalos cronolóxicos interrompidos non resolven por completo o problema da avaliación: se acontece que a introdución do programa en cuestión coincide con outros eventos, ou coa introdución doutros programas que teñan un impacto nos resultados, será imposible probar que o programa en cuestión causou o cambio que se observou. No caso dos programas que teñen un impacto retardado ou gradual, a interrupción na serie de períodos terá lugar un tempo despois de que se introduza o programa. A non ser que isto se teña en conta de antemán, pode ocorrer que o impacto do programa pase desapercibido.

Como pasa co método básico de antes e despois, os modelos baseados en intervalos cronolóxicos interrompidos resultan máis convincentes no caso de programas cun impacto razoablemente grande, pois será relativamente fácil detectar o devandito impacto entre o 'ruído ambiente'. Isto significa que este método *non* sería adecuado para programas voluntarios, especialmente aqueles cun baixo número de participantes.

Un tipo particular de método de intervalos cronolóxicos interrompidos é o método de tratamento 'suprimido'. En circunstancias excepcionais introdúcese un programa que despois se suprime. Neses casos, esperaríase que as series cronolóxicas revelasen dúas 'interrupcións': unha no momento en que se introduce o programa e outra cando se suprime. Agardaríase que en gran medida o primeiro cambio quedase revertido após a supresión do programa. En principio, este modelo podería ser moi puxante; porén, úsase poucas veces ou ningunha nas avaliacións dos programas gobernamentais pola simple razón de que, cando se abandona unha política, non adoita haber demasiado interese por practicarlle a autopsia.

En último lugar, nos métodos de tratamento demorado ensáiase un programa nunhas poucas zonas antes de introducilo na nación enteira. Se os datos solicitados nos intervalos cronolóxicos mostran unha diferenza entre antes e despois nas zonas de ensaio, e despois esa diferenza se reproduce noutras zonas cando se implanta o programa a escala nacional, obtéñense probas moi firmes do impacto do programa.

Cando se traballa con métodos de intervalos cronolóxicos interrompidos é importante asegurarse de que as variables que se están medindo permanezan estables no tempo: cando as definicións e as prácticas de cómputo varían frecuentemente, é moito máis difícil –e ás veces imposible– utilizar eses datos como medidas válidas ou atribuírlle un motivo a unha interrupción na serie de intervalos.

2.6.3.3. Melloras de dobre diferenza

Tanto o método básico de antes e despois como o método dos intervalos cronolóxicos interrompidos computan a adicionalidade atribuíble a un programa calculando a diferenza (axustada) nos resultados entre os períodos anteriores e posteriores. Con todo, ambos métodos poden ser pouco convincentes se se dá a sospeita de que a diferenza pode deberse a outros eventos que tivesen lugar máis ou menos ao mesmo tempo que se implantou o programa. Este problema pode afrontarse mediante os métodos de dobre diferenza.

Este tipo de método funciona comparando unha estimación de antes e despois referente aos participantes cunha estimación de antes e despois referente aos non participantes, e entendendo a diferenza como adicionalidade. A presuposición de identificación é máis verosímil ca no caso do indicador de antes e despois. Especificamente, pártese da base de que a media do cambio na medición do resultado alleo ao programa será a mesma para os participantes e os non participantes. Na práctica, isto significa que o indicador de dobre diferenza pode asimilar os cambios macroeconómicos ou os cambios no ciclo vital, a condición de que estes afecten de xeito similar a participantes e non participantes. Isto reafirma a necesidade de seleccionar un grupo adecuado de non participantes. A elección do grupo de comparación xustifícase frecuentemente porque ten tendencias similares ás do grupo experimental, no que respecta á variable nos resultados en cuestión, ao longo dun período prolongado antes da implantación do programa. Aínda que isto resulta tranquilizador, hai que ter en conta que é habitual axustar as estimacións de dobre diferenza para destacar as características observables; por tanto, a tendencia que hai que observar é a dos resultados regresivos, non a das propias medidas de resultados.

Ademais dos efectos individuais e transitorios que caracterizan o indicador de antes e despois, tamén se ten en conta un efecto común aos individuos pero que varía no tempo (efecto de tendencia). Como xa se fixo notar, o indicador de antes e despois elimina os efectos individuais. A vantaxe do indicador de dobre diferenza é que tamén elimina os efectos de tendencia. Así pois, o único efecto que permanece é o específico do individuo que varía no tempo: este non pode ser controlado e, de influír na decisión de participar no programa, a presuposición identificativa quedaría en cuestión e as estimacións resultadas estarían nesgadas (Heckman e Smith, 1999).

Aínda máis, tanto o indicador de antes e despois como o de dobre diferenza dependen de que a composición do grupo experimental non sufra mudanzas no período posterior ao programa. Se isto non se cumpre, a diferenza entre a verdadeira hipótese e o resultado previo ao programa non dará necesariamente unha media de cero en todos os individuos. Un

cambio así na composición é máis probable na repetición dos datos dun corte transversal, pero tamén pode darse con datos lonxitudinais se a mostra se fose reducindo co paso do tempo de xeito sistemático.

Alén dos grupos de control non equivalentes sobre os que se falou antes, é posible utilizar variables non equivalentes de resultados como modelos de dobre diferenza. Este método compara o cambio no tempo nos resultados baixo exame cos cambios noutros resultados relacionados. Estas variables non equivalentes de resultados teñen que satisfacer o requisito de cambiar historicamente en paralelo aos resultados de interese, pero non lles afecta a política que se está examinando.

Un exemplo famoso deste método foi a avaliación das probas obrigatorias de alcoholemia no Reino Unido en 1967. Comparáronse dúas variables de resultados: os accidentes nas noites do fin de semana, que se esperaba que variasen de frecuencia coa implantación das probas, e os accidentes nas horas puntas laborais, para os que non se prevían mudanzas canto á frecuencia.. O que se observou foi unha brusca caída na primeira variable e poucos cambios ou ningún na segunda: isto deu probas sólidas de que as probas tiveran un impacto. De novo, a principal razón pola que non adoita utilizarse este método é a dificultade de atopar resultados pertinentes cos cales comparar a principal variable de resultados.

2.6.3.4. Métodos de comparación de dous grupos equivalentes

Os modelos de avaliación baseados nos intervalos cronolóxicos son menos convincentes no caso de programas voluntarios con baixo número de participantes. Por este motivo deseñáronse métodos *quasi-experimentais* que funcionan precisamente nesas condicións, e poden ademais afrontar algunhas das limitacións dos métodos de dobre diferenza. Estes métodos son fundamentalmente dous: o método de comparación de dous grupos equivalentes que se describe aquí e outro modelo máis xeral que pode ser denominado 'modelado estatístico de datos xa existentes', dado que engloba varios métodos de análise no canto de ser un modelo alternativo *quasi-experimental* como tal. O primeiro destes dous modelos utilízase normalmente cando hai que recompilar datos sobre os resultados por medio dunha sondaxe, caso no cal o tamaño dos grupos de participantes e non participantes terá que ser razoablemente pequeno. Os métodos de modelado 'estatístico' utilízanse cando hai dispoñibles datos sobre os resultados, procedentes polo xeral dos rexistros administrativos ou de sondaxes elaboradas para outro propósito.

O método de comparación de dous grupos equivalentes fai uso do feito de que, nos programas voluntarios, a adicionalidade se concentra nos participantes no programa en vez de estar distribuída por toda a poboación potencialmente afectada. De feito, dáse por sentado que o programa non ten impacto ningún sobre os non participantes, de maneira que deste grupo se pode extraer unha estimación hipotética razoable de cales serían os resultados en ausencia do programa. Igual que con todos os métodos quasi-experimentais, compáranse dous grupos: un grupo de programa e un grupo de comparación. No método de comparación de dous grupos equivalentes, o grupo de programa selecciónase de entre os participantes e o grupo de comparación de entre os non participantes, e compílanse datos achega dos resultados dos membros dos dous grupos. Se se considera que os datos de resultados son caros de recompilar –o cal sucedería se houbese que facelo por medio dunha sondaxe–, escollerase un grupo de comparación do mesmo tamaño que o grupo de programa (de aí que a comparación sexa entre dous grupos *equivalentes*), aínda que tamén se poden comparar grupos non equivalentes.

O método da equivalencia dá por sentado que a selección pode explicarse unicamente mediante as características observables. En principio, a aplicación do método é sinxela: por cada individuo do grupo de programa, búscase un individuo equivalente no grupo de comparación. A elección deste emparellamento vén ditada polas características observables. O que se necesita é casar os individuos do grupo de programa con cadanseu individuo de características similares; pódese entón calcular o efecto medio do tratamento como a diferenza media en resultados entre un grupo e o outro.

Os métodos de comparación de dous grupos equivalentes son controvertidos pola forma de selección do grupo de comparación. Para que un grupo de comparación proporcione unha estimación hipotética axustada das condicións caso de non existir o programa, o grupo debe ter o mesmo perfil que o grupo de programa no referente aos resultados, coa única excepción de que non está exposto ao programa (ou, neste caso, *decide* non estalo). Pero, ao revés ca no caso dos programas obrigatorios, neste caso o grupo de programa escóllese a si mesmo e, en gran medida, as súas razóns para participar non se coñecerán ou comprenderán por completo; en consecuencia, resulta moi difícil construír un grupo de comparación axeitado.

Daquela, para evitar un nesgo no método de comparación de dous grupos equivalentes, deben satisfacerse os dous criterios seguintes:

- os avaliadores deben coñecer exhaustivamente os factores que inflúen na participación nunha política
- deben existir datos sobre estes factores para todos os participantes e non participantes, ou para unha parte de ambos

Se estes dous criterios se cumpren, a comparación entre dous grupos equivalentes é un método sólido de avaliación; en caso contrario, a equivalencia non será perfecta e daranse diferenzas residuais e incontroladas entre o grupo de programa e o de comparación. Estas diferenzas introducirán un nesgo nas estimacións de adicionalidade.

2.6.3.5. Modelado estatístico de datos xa existentes

O método de comparación de dous grupos equivalentes crea un grupo de comparación aproximadamente do mesmo tamaño ca o grupo de programa, pero seleccionado de entre un colectivo moito maior: o de non participantes. Isto é adecuado se a recompilación de datos é cara, dado que nese caso a selección dun grupo de comparación maior do estritamente necesario é un dispendio. Porén, se poden recompilarse datos cun custo baixo ou sen custo ningún, pódese obter unha maior precisión usando un grupo maior de non participantes. A bibliografía sobre econometría suxire varios xeitos de estimar hipoteticamente os efectos da non existencia do programa baixo estas circunstancias, xeitos que son todos de natureza técnica e inclúen: o cotexo de puntuacións de propensión por medio de avaliación *kernel,* o indicador de variables instrumentais e o indicador de selección Heckman. Todos estes métodos necesitan presuposicións moi importantes para unha estimación hipotética obxectiva e tenden a caer nos mesmos problemas de interpretación ca o método estándar de comparación de dous grupos equivalentes.

2.6.3.6. Métodos de comparación de áreas equivalentes

A comparación de áreas equivalentes soamente pode utilizarse para medir a adicionalidade cando se ensaia unha política en varias zonas xeográficas: a forma básica da comparación de áreas equivalentes parte dun pequeno número de zonas en que se proba a nova intervención, de entre as cales se selecciona o grupo de programa. Seguidamente emparéllanse as zonas con outras áreas de características similares onde non se estea probando a intervención e selecciónase de entre elas o grupo de comparación. Nunha variante do método básico, selecciónase o grupo de comparación de entre o resto do país, non de zonas equivalentes. A continuación recompílanse os resultados de ambos os grupos.

A interpretación dos métodos de comparación de áreas equivalentes pode resultar problemática, porque, malia que as diferenzas observadas entre o grupo de programa e o de intervención poden atribuírse á intervención da política, tamén poderían imputarse a diferenzas no perfil dos dous grupos ou a outras características locais que non se tiveron en conta durante o proceso de emparellamento. A primeira diferenza é a que desexamos medir; as outras son un problema potencial, pero ata certo punto poden abordarse buscando diferenzas entre os grupos no estadio de análise, por exemplo, por medio de análises regresivas ou medindo as mostras con respecto a un perfil común.

As comparacións de áreas equivalentes son candidatos naturais a un modelo baseado en diferenzas e similitudes. Se poden recompilarse datos de varios intervalos cronolóxicos sobre os resultados para a zona piloto e a de control e se pode probar que estas dúas series de intervalos discorreran en paralelo –durante un tempo, polo menos– antes de que se introducise o programa nas zonas piloto, tomar a diferenza entre antes e despois nas dúas zonas e comparalas pode proporcionar un indicador do efecto do programa moito mellor ca o que se obtería mediante a simple diferenza entre as dúas zonas. Este modelo é especialmente útil se o emparellamento das zonas é só aproximado, dado que nese caso pode haber diferenzas previas importantes entre as dúas zonas.

2.7. Análises económicas

As análises económicas provén dunha visión particular da avaliación. Estes métodos relacionan o impacto con outros estadios dun ciclo de intervención, principalmente as achegas e o obxectivo da política, para medir a eficacia e a eficiencia. A eficacia compara o que se fixo co que estaba previsto orixinalmente, isto é, compara os resultados e/ou impactos reais cos agardados, mentres que a eficiencia examina a relación entre os resultados e/ou impactos cos recursos (principalmente os recursos financeiros) empregadas para conseguilos. As análises económicas inclúen as sondaxes de control do gasto público e técnicas máis sofisticadas, como análises custo-beneficio, custo-eficacia e custo-utilidade.

2.7.1. Sondaxes de control do gasto público

As sondaxes de control do gasto público seguen o fluxo de fondos públicos e determinan ata que punto os recursos chegan realmente aos grupos aos que están destinados. As sondaxes examinan a maneira, a cantidade e o tempo en que se lles conceden recursos aos diversos niveis do goberno, especialmente ás unidades responsables da prestación de servizos sociais como a sanidade ou a educación. As sondaxes de control do gasto público poden simplemente comparar os custos e gastos programados de diferentes iniciativas sen ter en conta os resultados pretendidos ou conseguidos. As limitacións de semellantes valoracións e avalia-

cións resultan evidentes –posto que nos indican moi pouco sobre a *eficacia relativa ou os efectos beneficiosos* das diversas intervencións– e teñen, de seu, moi pouco valor para a avaliación de políticas. Con todo, moitas veces estas sondaxes lévanse a cabo como parte doutras máis amplas sobre a prestación de servizos, que se centran na calidade do servizo, nas características da súa prestación, na súa xestión ou nas súas estruturas incentivas, e que contribúen a fomentar a responsabilidade cando hai pouca información financeira dispoñible.

2.7.2. Análise da relación entre custo e beneficio, entre custo e eficacia e entre custo e utilidade

Outros tipos de valoración e avaliación económica, máis puxantes e máis útiles para a creación de políticas, inclúen as análises das relacións custo-eficacia e custo-beneficio, ferramentas para xulgar se os resultados e impactos dunha actividade poden xustificar os seus custos. A análise da relación custo-beneficio mide tanto as achegas como os resultados en termos monetarios (World Bank, 2004). A análise da relación custo-eficacia mide as achegas en termos monetarios e os resultados en termos cuantitativos non monetarios (como as melloras na alfabetización dos escolares). Ambas as técnicas poden contribuír a saber que proxectos dan máis resultado a cambio do investimento, e tamén demostraron ser ferramentas útiles para convencer a quen deseñan as políticas e controlan os recursos de que os beneficios dunha intervención a xustifican, aínda que na práctica é frecuente que non se dispoña de datos necesarios para os cálculos da relación custo-beneficio e os resultados previstos dependan en gran medida das presuposicións adoptadas. A análise custo-beneficio adoita implicar a consideración de usos alternativos dun recurso determinado ou o *custo en oportunidades* de facer algo fronte a facer outra cousa. Outro tipo de valoración económica é a análise da relación entre custo e utilidade, que avalía a utilidade de diversos resultados para diversos usuarios ou consumidores dunha política ou servizo. Normalmente, este método implica valoracións e avaliacións subxectivas de resultados utilizando datos cualitativos e cuantitativos.

Estas análises usan diversas ferramentas para estimar os custos e os beneficios das políticas ao longo do tempo, como a *taxa de desconto* para axustar o valor dos resultados que terán lugar no futuro.

2.8. Avaliacións cuantitativas e cualitativas

Unha última distinción dentro dos tipos de avaliación refírese ao carácter dos datos recollidos e á maneira en que se analizan. Neste sentido é habitual distinguir entre avaliacións cuantitativas e cualitativas. As avaliacións cualitativas están pensadas para "permitir ao avaliador estudar determinados temas en profundidade e detalle" (Patton, 2002). Normalmente a profundidade e o detalle son imprescindibles para decidir que preguntas formular nunha avaliación e para identificar as condicións situacionais e contextuais baixo as cales funciona ou deixa de funcionar unha política, un programa ou un proxecto. Os métodos cualitativos de avaliación son especialmente importantes para as avaliacións formativas, as cales, como apunta unha vez máis Patton (2002), adoitan limitarse por completo a un contexto específico. Sucede a miúdo que nas avaliacións cualitativas non hai intentos de xeneralizar os achados máis alá do contexto en que un está traballando.

En cambio, as avaliacións cuantitativas poden usar técnicas econométricas como as xa mencionadas neste capítulo para medir o impacto ou poden resumir grandes cantidades de datos compilados con sondaxes feitas ex profeso.

Elliot Eisner argumentou que, aínda que as técnicas cuantitativas poden proporcionar algunha información útil, "a avaliación require un mapa interpretativo sofisticado, non só para separar o trivial do significativo senón tamén para entender o significado daquilo que se coñece" (Eisner, 1994). Con todo, en última instancia, a decisión de adoptar métodos cuantitativos ou cualitativos depende dos fins da avaliación e da natureza da intervención que se avalía, aínda que tamén teñen importancia as tendencias xerais do clima político. No contexto da avaliación de intervencións de desenvolvemento, White (2005) apunta que o abandono do crecemento como medida de desenvolvemento na década de 1970 se reflectiu nunha modificación dos xeitos de calcular a eficacia. Esta modificación obedeceu en parte á opinión de que os sectores sociais eran menos receptivos ás análises económicas da relación entre custo e beneficio e a un desexo de centrarse directamente en resultados non económicos, como a igualdade de xénero. Pensábase que a análise da relación entre custo e beneficio non podía aprehender estes asuntos e que se faría necesario, xa que logo, un modelo máis cualitativo. Para a década de 1980, os modelos cualitativos pasaran a dominar os estudos de avaliación que se facían para as axencias de desenvolvemento, mudanza que nese momento se viu reforzada pola importancia que se lle daba ao proceso.

Sen dúbida estes son puntos importantes que poden pasarse por alto nun estudo estritamente económico; e adoita ser xusto dicir que os 'proxectos de proceso', cuxo fin principal adoita ser o desenvolvemento institucional, están habitualmente demasiado distantes dos resultados finais de desenvolvemento como para que poida cuantificarse o seu impacto nestes últimos. No entanto, a nova centralidade concedida aos resultados no contexto de iniciativas como os obxectivos de desenvolvemento do milenio das Nacións Unidas está facendo que os creadores das políticas estean volvendo prestar atención ás avaliacións cuantitativas. Os métodos cualitativos teñen dificultades para responder preguntas cruciais para os responsables das políticas como: ata que punto están as intervencións das axencias traendo progreso nos eidos relacionados coas metas de desenvolvemento do milenio?

Na próxima sección centrarémonos nos métodos de compilación de datos.

3. MÉTODOS DE COMPILACIÓN DE DATOS

3.1. Introdución

Os diferentes tipos de avaliación que se presentaron poden acometerse por medio do uso de varios métodos de compilación de datos, que se presentan nesta sección. Aínda que algunhas avaliacións utilizan un único método para a compilación de datos, moitas empregan unha mestura de métodos, combinando os datos de varios instrumentos e enfoques. Neses casos, ponse especial coidado en evitar que os datos se perdan ou dupliquen como consecuencia de ter que combinar datos de diferentes fontes. Os métodos primarios e secundarios que se presentan aquí son:

- informe documental
- sondaxe formal
- entrevistas
- grupos de discusión e outras formas de consulta
- monografías
- observación de participantes
- métodos participativos[5]

3.2. Revisión documental

A maior parte das avaliacións tenden a facer uso da revisión de documentos. Nalgúns casos, os documentos son a fonte de 'reconstrución' dunha base para a avaliación, dado que rexistran o que un proxecto pensaba conseguir, as condicións desde as cales se desenvolveu, como se desenvolveu, que cambios se levaron a cabo e como se levaron á práctica. Noutros casos, a análise documental acompaña a outros métodos de compilación de datos, por exemplo entrevistas en profundidade ou sondaxes.

As revisións documentais teñen grandes vantaxes: ademais da precisión engadida que proporcionan os datos documentais e administrativos, está tamén a vantaxe de que os datos non adoitan estar suxeitos aos nesgos que sofren as sondaxes debido aos casos en que non hai resposta, e o custo da compilación de datos é menor. Non obstante, o avaliador ten menos liberdade para determinar o contido dos datos á súa disposición.

Cada vez utilízanse máis as revisións sistemáticas da bibliografía de investigación existente como método válido e fiable de utilizar a evidencia xerada pola investigación; tamén poden permitir o establecemento dunha visión acumulativa de devandita evidencia. As revisións sistemáticas difiren doutros tipos de síntese da investigación (por exemplo, das revisións narrativas e do reconto de votos) en que son máis rigorosas á hora de buscar e atopar evidencias, dado que teñen criterios explícitos e transparentes para valorar a calidade da evidencia xerada pola investigación, e especialmente para identificar e controlar os diversos tipos de nesgo nos estudos existentes, en canto teñen xeitos explícitos de establecer a posibilidade ou imposibilidade de comparar distintos estudos e, en consecuencia, de combinar e establecer unha visión acumulativa do que nos di a evidencia actualmente existente.

Dous métodos comúns de sintetizar a evidencia actualmente existente son as revisións narrativas e o reconto de votos. As revisións narrativas tratan de identificar a bibliografía dispoñible acerca dun tema, coas súas metodoloxías, os seus achados e as súas limitacións. Poden presentar un resumo da investigación sobre un tema ou non. Tenden a identificar o alcance e a diversidade da bibliografía dispoñible, da cal moita será inconsistente ou non concluínte. Unha das súas grandes limitacións é que son case sempre selectivos. Non sempre implican unha busca sistemática de toda a bibliografía relevante usando fontes electrónicas e impresas, ademais de pescudando directamente para atopar estudos inéditos ou traballos en curso de elaboración: isto significa que as revisións narrativas tradicionais sobre a bibliografía sofren frecuentemente nesgos de selección ou de publicación. As revisións sistemáticas disenten tamén das narrativas en que explicitan os criterios de procura para identificar a bibliografía dispoñible e os procedementos polos que se valora e interpreta criticamente a devandita bibliografía. Isto proporciona un grao de transparencia por medio do cal os lectores poden determinar que evidencia se estudou e como se interpretou e se presentou.

Os recontos de votos tratan de acumular os resultados dun conxunto de estudos relevantes contando "cantos resultados son estatisticamente significativos nun sentido, cantos son neutrais (é dicir, que 'non teñen efecto') e cantos son estatisticamente significativos no outro sentido" (Cook *et al.,* 1992: 4). Considérase que a categoría máis numerosa, é dicir, a que teña máis votos, é a que representa os achados típicos ou modais, indicando pois o medio de intervención máis eficaz. Un problema obvio dos recontos de votos é que non teñen en conta o feito de que hai estudos que son metodoloxicamente superiores a outros e, xa que logo, merecen ter máis peso. As revisións sistemáticas da literatura distinguen entre estudos

que abordan mostras de maior ou menor tamaño, estudos de maior ou menor poder e precisión, e valóranos como corresponda.

Un tipo especial de revisión sistemática é a 'metaanálise', que agrega os resultados dos estudos comparables e "combina os efectos de tratamento calculado por cada estudo nun efecto de tratamento do conxunto de todos os estudos" (Morton, 1999). Isto, porén, non sempre é posible, posto que só pode ocorrer se hai unha consistencia real entre os estudos primarios en canto aos tipos de intervención que analizan, á poboación que estudan e aos resultados que valoran (véxase Deeks, Altman e Bradburn, 2001).

Slavin (1984, 1986) aventurou que o método que se use para xerar evidencia investigadora é menos importante que a calidade dos estudos primarios emprendidos, usen as opcións metodolóxicas que usen. Suxire que o que se necesita é a 'síntese da mellor evidencia', na que "os avaliadores apliquen criterios de inclusión consistentes, xustificados e claramente expostos *a priori*", aos traballos que se examinan. Para Slavin, os estudos primarios deben ser "relevantes para o tema en cuestión, estar baseados nun esquema que minimice o nesgo e ter validez externa".

Xeralmente os resultados das revisións documentais analízanse por medio da análise teórica (onde se estudan os documentos para saber se obedecen a unha teoría ou explicación predefinida), a análise estrutural (onde se estuda en detalle a estrutura do documento, é dicir, como está construído, en que contexto se sitúa, como transmite as súas mensaxes) e a análise de contidos (onde se estuda e se compara a información de determinados tipos de documento).

3.3. Sondaxes formais

As sondaxes formais son un método para compilar información estandarizada a partir dunha mostra seleccionada de individuos e organizacións. A miúdo, as sondaxes compilan información comparable para un número de casos relativamente grande e proporcionan datos de base cos que comparar o rendemento dunha estratexia, dun proxecto ou dun programa. Así pois, as sondaxes poden ser unha fonte valiosa para unha avaliación formal do impacto dun programa ou proxecto. Hai diversos tipos de instrumentos para as sondaxes que poden utilizarse para recompilar a información necesaria para responder as preguntas da investigación; entre estes, os seguintes:

3.3.1. Cuestionarios

Os cuestionarios compilan información por medio de preguntas prefixadas. O cuestionario pode formulalo un entrevistador (cara a cara ou por teléfono) ou pode completalo o entrevistado (no caso de sondaxes por correo ou en liña). Os cuestionarios poden recoller información fáctica e información referente aos comportamentos e ás actitudes, ademais de medir os coñecementos dos entrevistados, aínda que estes últimos só poden recollerse de maneira fiable se un entrevistador formula o cuestionario ou se o entrevistado o cobre nun contorno controlado. A forma en que se recompilen os datos pode influír na fiabilidade e precisión da información obtida. Por exemplo, a precisión da información sobre comportamentos 'problemáticos', como o consumo de drogas, pode diferir dependendo de se os datos os recolle un entrevistador ou se o propio entrevistado enche un formulario.

3.3.2. Diarios

Os diarios posibilitan a compilación prospectiva de información, é dicir, a medida que se produce un evento. Son unha forma de sondaxe para cubrir polo entrevistado, de quen se require que rexistre detalles do comportamento de interese durante un período específico. Espérase capturar así os detalles do comportamento habitual dos entrevistados. Os diarios poden capturar información sobre o comportamento moito máis detallada do que é posible decote noutros tipos de sondaxe, e pode usarse ao mesmo tempo que os cuestionarios estruturados.

3.3.3. Medicións

Poden efectuarse medicións para recompilar información fáctica como a estatura dos entrevistados, o seu peso, a súa tensión vascular, os seus niveis de ferro en sangue, etc. Igual ca no caso dos diarios, estas medicións poden efectuarse en conxunción coa información obtida dun cuestionario (ou un diario). Cómpre desenvolver protocolos para asegurarse de que as medicións se efectúan de xeito estandarizado. Pode ser necesaria a aprobación ética.

3.3.4. Probas

Como parte do proceso de entrevistas da sondaxe, poden administrarse probas para medir a capacidade dos entrevistados para desempeñar determinadas tarefas, como ler ou camiñar. Con frecuencia estas probas son ferramentas estandarizadas de valoración que se desenvolveron para un contexto determinado, como a valoración clínica ou educacional nun hospital ou unha escola. Como no caso da compilación de medicións, é necesario desenvolver protocolos que garantan que as probas se administran de xeito coherente e que poidan ser formulados (de modo fiable) durante unha entrevista correspondente a unha sondaxe.

3.3.5. Observacións

Poden levarse a cabo observacións de información fáctica, como o estado da vivenda do entrevistado. Os observadores deben recibir unha coidadosa formación para rexistrar a información de xeito coherente. Os datos observacionais poden compilarse ao mesmo tempo que outros tipos de información para obter unha imaxe máis detallada das circunstancias do entrevistado. A elección de instrumento de compilación de datos verase influída pola natureza das preguntas, o tipo de información que se require, o grao de detalle necesario, o grao requirido de precisión dos datos, as características da poboación da que se solicita a información, o tempo e o diñeiro.

As sondaxes poden ser transversais, se recompilan información sobre a poboación concernida nun momento determinado, ou lonxitudinais, se reúnen información sobre individuos determinados ao longo do tempo. Ademais, as sondaxes transversais poden dividirse en sondaxes *ad hoc* –feitas unha única vez–, continuas –nos que o traballo de campo ten lugar, por exemplo, cada mes do ano, sendo a mostra de cada mes representativa da poboación concernida– e repetidas –que teñen lugar en momentos regulares e determinados, por exemplo cada ano ou cada dous, co traballo de campo concentrado nuns cantos meses, o cal permite medir o cambio global no nivel agregado, dado que as estimacións dunha sondaxe poden compararse con outras da mesma serie pero non permite saber se o cambio observado tivo lugar gradualmente ou non, cousa que si permiten as sondaxes continuas–. Igual ca coas continuas, non hai nada no modelo das sondaxes repetidas que requira un solapamen-

to das mostras en diferentes momentos. Isto distíngueas doutros tipos de sondaxe, como as de mostraxe rotatoria ou os estudos lonxitudinais sen rotación. As enquisas de mostraxe rotatoria prográmanse a intervalos regulares ou continuamente e utilizan mostraxes rotatorias, é dicir, inclúese certo número de persoas na sondaxe, examínanse algunhas veces e despois exclúense da enquisa: non se intenta seguir os entrevistados ou unidades de mostraxe nin tampouco relacionar os resultados duns ou doutros a través do tempo para obter estimacións lonxitudinais. Nos estudos lonxitudinais sen rotación, séguese a través do tempo os participantes para crear un rexistro lonxitudinal; con todo, estes datos diacrónicos non poden extrapolarse á poboación xeral.

3.4. Entrevistas

As entrevistas en profundidade son probablemente a forma máis habitual de investigación cualitativa na avaliación. Crese que os informes persoais orais teñen unha importancia central na investigación social debido ao seu poder de elucidar o significado (Hammersley e Atkinson, 1995). Os informes individuais e persoais presentan a linguaxe que a xente utiliza e os aspectos que destacan, e permiten que a xente dea explicacións claras acerca das súas accións e decisións.

As entrevistas en profundidade dan a oportunidade de compilar datos ricos e detallados, dado que o entrevistador pode 'espremer' o tema e incitar o entrevistado a que profunde máis e máis nas súas respostas. Son perfectas para a exploración en profundidade dun tema que lle proporcione ao investigador unha visión detallada do mundo do entrevistado, as súas crenzas, experiencias e sentimentos, e as explicacións que dá das súas convicción ou accións. Estas entrevistas tamén se prestan ben a explorar procesos complexos ou a desentrañar a toma de decisións. Un bo entrevistador establece unha comunicación co seu entrevistado, o cal facilita a exploración de temas conflitivos, dolorosos ou problemáticos.

O grao de estruturación das entrevistas varía: un trazo clave das entrevistas cualitativas é que as preguntas non están determinadas de antemán, senón que o entrevistado ten certa influencia sobre a dirección e a cobertura da entrevista. Nalgúns estudos, quizais particularmente naqueles cuxo propósito é revelarlle ao investigador un mundo social co que non está familiarizado, a entrevista pode estar relativamente desestruturada, de maneira que o entrevistador formula preguntas moi amplas e o entrevistado moldea o resultado. Noutros casos, o investigador terá unha consciencia máis clara dos temas a explorar e xogará un papel máis activo na dirección dos temas da entrevista. Por veces, os termos 'desestruturado' e 'semiestruturado' denotan diversos graos nos que a investigación dirixe a entrevista, aínda que non sempre se utilizan de xeito coherente.

Nos modelos biográficos adoita usarse un tipo especial de entrevista, historias vitais e narracións onde os investigadores retornan con frecuencia a un informante para obter máis datos ou manter máis entrevistas. Por exemplo, poden estudar a perspectiva dunha familia entrevistando varios dos seus integrantes, ou a dos membros dunha comunidade específica.

3.5. Grupos de discusión e outras formas de consulta

Os grupos de discusións ou discusións en grupo consisten habitualmente en ata dez persoas reunidas para falar dun tema ou de varios. O grupo modérao ou facilítao un investigador. Aínda que os grupos de interese adquiriron unha imaxe un tanto turbia, son un método rigoroso e ben establecido de investigación e avaliación social. Nos grupos de discusión, os

datos son moldeados e matizados por medio da interacción do grupo: escoitar a participación doutros estimula o pensamento e anima a xente a reflexionar sobre as súas propias opinións ou sobre o seu comportamento, xerando novo material.

Os grupos de discusión poden funcionar moi ben cando se trata de temas abstractos ou conceptuais, que nunha entrevista poderían deixar o entrevistado en branco. Tamén poden ser utilizados para temas conflitivos, a condición de que as características sociais dos participantes e a súa conexión co tema da investigación sexan similares abondo como para crear un contorno que transmita sensación de seguridade.

Outras formas de consulta máis innovadoras son as seguintes:

- O método Delphi (Adler e Ziglio, 1996; Cantrill *et al.,* 1996; Crichter e Gladstone, 1998). Este é un proceso iterativo especialmente orientado á predición. Pídeselle a un grupo de expertos que respondan individualmente unha serie de preguntas, ben por medio dunha enquisa ben usando investigación cualitativa. A seguir, difúndense as respostas entre os membros do panel, a quen se lles pide que valoren as súas propias respostas, que despois se difunden e se matizan sucesivamente ata chegar a un consenso ou a unha disensión acordada. O grupo non se reúne fisicamente.
- A técnica de grupo nominal é unha variante do método Delphi que segue un patrón similar para obter as primeiras respostas dos membros do panel. Con todo, tras esa primeira fase lévanse a cabo novas iteracións usando un formato relativamente similar a un grupo de discusión. O obxectivo do grupo é acadar un consenso nas áreas de acordo e desacordo.
- Os xurados cidadáns (Coote e Lenaghan, 1997; Davies *et al.,* 1998; White *et al.,* 1999). Reúnese un grupo de entre doce e vinte persoas durante varios días para escoitar varias 'testemuñas' expertas e formularlles preguntas, deliberar e debater entre elas, e facer recomendacións sobre as accións que se deben tomar, que poden estar consensuadas ou non.
- Sondaxes deliberativas (Fishkin, 1995; Park *et al.,* 1999). Trátase dun conxunto de actividades dirixidas a explorar como muda a opinión pública cando o público ten a oportunidade de informarse en profundidade sobre un tema. Lévase a cabo unha enquisa para establecer un punto de referencia da opinión pública. Os participantes asisten a un evento conxunto, polo xeral durante unha fin de semana, que inclúe debates en grupos pequenos, conferencias de expertos e sesións políticas nas que voceiros dos partidos responden preguntas. Repítese a enquisa para medir en que sentido e en que medida cambiou a opinión.
- Congresos ou obradoiros de consenso (Seargeant e Steele, 1998). Neste modelo, un panel dunhas quince persoas trata de definir as cuestións que desexa afrontar con respecto a un tema determinado, consulta expertos, recibe información, delibera e trata de conseguir un consenso. O panel elabora o seu propio informe e preséntao nun congreso aberto, onde se debate de novo.
- 'Valoración participativa'. Historicamente, este método usouse no traballo para o desenvolvemento das colonias, pero agora é máis frecuente no Reino Unido. Está pensado para involucrar a xente, especialmente a de comunidades socialmente excluídas, en decisións que incumben ás súas vidas. Combina varias ferramentas visuais (mapas) con discusións en grupo e entrevistas semiestruturadas.
- 'Planificación de verdade'. É un proceso de consulta da comunidade onde se usan modelos para animar os residentes a explorar e priorizar opcións para a acción (Gibson, 1998).

3.6. Monografías

Os estudos monográficos úsanse para compilar datos descritivos por medio do exame intensivo dun fenómeno nun individuo, grupo ou situación. Os modelos monográficos úsanse cando se necesita unha comprensión en profundidade moi detallada que sexa holística, integral e contextualizada. Permiten establecer comparacións entre diferentes suxeitos dentro do mesmo caso, entre casos e entre grupos de diferentes casos. Así, por exemplo, no contexto dunha investigación baseada nas escolas, podería estudarse como diferentes persoas da mesma escola teñen distintas interpretacións dunha nova iniciativa educativa, como diferentes escolas a levaron á práctica ou como perciben a iniciativa os directores en contraste cos docentes ou os alumnos. As monografías son intensivas e, xa que logo, poden ser caras ou complexas ou esixir moito tempo, pero poden proporcionar unha comprensión moi profunda á avaliación de políticas. Os datos observacionais e etnográficos poden ser triangulados por outra xente que participe na actividade observada ou por outros investigadores (especialmente onde se utilizaron gravacións de vídeo ou audio).

Por tanto, os estudos monográficos son especialmente útiles á hora de estudar fenómenos infrecuentes ou complexos. Robert Stake, entre outros, foi un firme partidario das monografías e de que o avaliador elabore unha 'descrición densa'. Pensa que os puntos de vista dos interesados son un elemento crucial das avaliacións e que os estudos monográficos son o mellor método tanto para representar as convicción e os valores das persoas interesadas como para presentar os resultados dunha avaliación.

Stake argúe que existen realidades múltiples e que os puntos de vista dos interesados deben figurar na avaliación, pero tamén sostén que as persoas interesadas non participan na avaliación como quererían os teóricos da participación. Oponse á participación das persoas interesadas tal como se describe máis arriba e argumenta que a avaliación é tarefa do avaliador (Alkin, Hofsetter e Ai, 1998: 98), principalmente a través do traballo monográfico, que pode incluír a observación, a análise de documentos relativos a un caso particular, entrevistas en profundidade e outros métodos de recompilación de datos.

3.7. Observación de participantes

Un dos principais xeitos que ten a investigación social de entender unha actividade, un grupo ou un proceso é aproximarse o máis posible, sen chegar a alterar o seu funcionamento 'natural' (Hammersley e Atkinson, 1995). Nun extremo, isto pode facerse sendo un observador absolutamente desapegado dunha situación social, traballando do modo máis discreto posible, observando, escoitando e recordando detalles; no outro extremo, un pode unirse ao grupo ou actividade en cuestión e participar nel como membro para aprender desde dentro. Esta opción pode ou non implicar 'facerse indíxena', isto é, involucrarse tanto no grupo, a actividade ou o proceso que se perda a obxectividade e a condición esóxena. Evidentemente, entre estes dous extremos hai opcións: un pode traballar como observador-participante ou como participante-observador, sendo a diferenza o grao de desapego e implicación que pode ter o investigador social.

3.8. Métodos participativos

Os métodos participativos fan posible que as persoas afectadas por un proxecto, un programa ou unha estratexia se involucren activamente na toma de decisións, e creen un vínculo cos resultados e coas recomendacións da avaliación. Son unha ferramenta útil para identi-

ficar os problemas durante os procesos de implementación e para aprender acerca das condicións locais e das perspectivas e prioridades da poboación local, para así deseñar intervencións máis receptivas e sustentables, aínda que ás veces se consideran métodos menos obxectivos ca os que dependen exclusivamente de avaliadores externos e poden –se é que se van involucrar en profundidade no proceso de avaliación– consumir demasiado tempo dos axentes locais. O punto de partida da maior parte do traballo participativo e das valoracións sociais é a análise da poboación afectada, que se utiliza para desenvolver unha comprensión das relacións de poder, influencia e intereses dos diversos suxeitos involucrados nunha actividade, e para determinar quen debería participar na avaliación e cando. Outros modelos comúns dos métodos participativos son a 'valoración de beneficiarios' (que implica o contacto sistemático cos beneficiarios do proxecto para identificar e deseñar iniciativas de desenvolvemento, localizar obstáculos para a participación e subministrar información para mellorar os servizos e actividades) e o 'seguimento e avaliación participativa' (no que persoas interesadas a diferentes niveis traballan xuntas para identificar problemas, compilar e analizar información e xerar recomendacións).

Inversamente a outros autores, como Scriven ou Eisner, que consideran o avaliador un 'valorador', Guba e Lincoln (1989) achan que as persoas afectadas son quen establecen primordialmente o valor. Este punto de vista baséase na idea de que, lonxe de existir unha realidade única, hai realidades múltiples baseadas nas percepcións e interpretacións dos individuos a quen incumbe o programa obxecto de avaliación. Así, Guba e Lincoln coidan que o papel do avaliador é facilitar as negociacións entre individuos que reflictan esas realidades múltiples e avogan polos métodos participativos.

David Fetterman foi un paso máis alá en *Empowerment Evaluation* (Fetterman *et al.,* 1996), onde describe a avaliación como un proceso que fomenta a autodeterminación entre os participantes na avaliación do programa e que acotío comprende "formación, facilitación, propugnación, iluminación e liberación". O obxectivo desta avaliación, que confire poder aos participantes, é potenciar a autodeterminación en lugar da dependencia, cousa que se consegue facendo que, no esencial, os afectados polo programa efectúen as súas propias avaliacións. O avaliador externo serve moitas veces como conselleiro ou auxiliar adicional, facilitándolles aos participantes os coñecementos e as ferramentas necesarias para a avaliación continua e a atribución de responsabilidades. Para el, o punto final da avaliación non é a valoración do programa; o valor e a valía non son estáticos: considera a avaliación un proceso continuo que "pode desenvolverse para acomodar cambios na poboación, nos fins, nas valoracións e nas forzas externas" (Fetterman, 1998).

4. COMO SABER SE ALGO FUNCIONOU? DEFININDO A BOA AVALIACIÓN DUNHA POLÍTICA

A visión tradicional da avaliación de políticas, xurdida nos Estados Unidos na década de 1960 –un período caracterizado pola necesidade urxente de avaliar os programas da *Great Society* do goberno estadounidense e polo optimismo respecto do coñecemento científico tras un período ateigado de éxitos para as ciencias naturais–, baséase na idea de que é posible darlles respostas científicas (obxectivas) ás preguntas incluídas en calquera proceso de avaliación.

En tempos máis recentes esta concepción foi atacada desde varios puntos, algúns dos cales descritos xa neste artigo. O argumento básico da maioría das críticas á visión tradicional é que as presuposicións baixo as cales opera este modelo de evolución non se sosteñen.

A avaliación é unha empresa política e social, onde as diferenzas de valor son relevantes, non só unha tarefa técnica e científica. Hai, ademais, múltiples perspectivas: isto non se debe só á ambigüidade dos resultados das políticas, programas e iniciativas na práctica, senón tamén ás disensións sobre que tipo de criterios avaliativos son significativos ou xustos nunha situación determinada (Majone, 1989). Subirats (1994) argúe que estas ambigüidades e estes problemas non poden resolverse empregando simplemente máis e mellores técnicas de medición. Os avaliadores 'pluralistas' argumentan que a avaliación debería utilizarse como un instrumento para crear confianza e consenso por medio de discusións conxuntas iterativas, non para pescudar que funcionou e que non. Así, Subirats (1994) afirma que o avaliador "non debe actuar en solitario, decidindo arbitrariamente se o programa do que se trata é bo ou malo; debe, máis ben, actuar como mediador entre as diversas opinións".

En certo sentido, este argumento central é unha proposición sorprendente. Aínda que é certo que poucos avaliadores sosterían que as premisas da avaliación tradicional son completamente válidas e menos aínda descoidarían a importancia de tratar cos distintos afectados no proceso de avaliación e de escoitalos, parece excesivo suxerir que unha avaliación realizada por un experto independente será indefectiblemente 'arbitraria' se a acomete en solitario –se se respectan, varios criterios como a solidez e a fiabilidade do proceso de avaliación poden controlar en gran medida a arbitrariedade– ou que a función primordial do avaliador é mediar entre diferentes afectados máis que elaborar unha valoración das políticas ou os programas, de acordo coas condicións contidas nos termos de referencia do proxecto.

Isto débese a varios motivos. En primeiro lugar, crer que a discusión sistemática pode conducir ao consenso pode resultar excesivamente optimista no contexto de moitas políticas. En segundo lugar, o marco temporal que requiriría as avaliacións que verdadeiramente aspirasen a este fin sería irrealizable, aínda supoñendo que fose posible. En terceiro lugar, agarda demasiado dos avaliadores. Existen, desde logo, outros casos en que o diálogo entre diferentes afectados no proceso da política é máis axeitado e disponse de maiores recursos para levar ao consenso ca durante un proceso de avaliación. Aínda que as avaliacións poden –e deben– valorar a relevancia das políticas, os programas e as iniciativas e dos debates que conduciron á adopción dunha política determinada (algo que decote pasan por alto os críticos pertencentes á tradición 'pluralista' da avaliación[6]), non poden absorber o proceso deliberativo da creación de políticas, porque esa función, tendo en conta os rápidos cambios de dirección das políticas no mundo actual, chegaría de todas as maneiras demasiado tarde con excesiva frecuencia. En derradeiro lugar, pasa por alto que os avaliadores son contratados para levar a cabo unhas tarefas especificadas nos termos de referencia do seu proxecto de avaliación; poucas veces poden permitirse o luxo de informar soamente sobre os diversos puntos de vista dos afectados e tratar de chegar a un consenso. Nun contexto de diminución dos fondos públicos e incremento do número de axencias que compiten por estes fondos, os creadores de políticas e o resto dos responsables necesitan valoracións do que funcionou e do que non para informar e lexitimar as súas demandas e mais as súas decisións. Se a avaliación se acomete con rigor e os creadores de políticas a toman en serio (cousa que fan cada vez máis, polo menos no caso de programas e intervencións de magnitude), o aspecto valorativo da avaliación é de grande importancia. Nel reside gran parte da utilidade da avaliación de políticas.

A elección entre distintos tipos de avaliación e distintos métodos de compilación de datos é, así pois, ata certo punto, unha cuestión técnica, aínda que condicionada polas limitacións dos recursos e polos contextos políticos. Malia que, como se salientou máis arriba, neste eido

existe unha gran variedade, hai varios criterios que poden usarse para identificar as boas avaliacións de políticas que poden ser de axuda para os responsables das decisións. As boas avaliacións caracterízanse por:

- Un conxunto definido de cuestións de investigación que son o bastante **específicas** como para poder ser levadas á práctica durante a investigación. As cuestións de investigación amplas ou imprecisas conducen con facilidade a estudos insatisfactorios que, simplemente, non proporcionan novos coñecementos. Isto pode impedirse ao principio do proceso avaliador dedicando máis tempo a definir que fai falta saber. No entanto, é frecuente que durante o proceso de investigación aparezan cuestións adicionais, o que significa que é posible que haxa que revisar e que emendar as cuestións iniciais á vista destes desenvolvementos.

- Ser **coherentes**. Debe haber coherencia entre as cuestións de investigación e a poboación obxecto de estudo; esta debe ser a poboación que vaia fornecer da información máis directa e profunda sobre o tema en cuestión. Como se comentou máis arriba, todos os afectados deben participar na avaliación, para mellorar a súa calidade e para facer os seus resultados máis interesantes para quen poden adoptar as súas recomendacións e producir un cambio.

- Ser **lóxicas**. Debe existir unha relación lóxica entre as cuestións de investigación e os métodos de compilación de datos utilizados –incluíndo a mostraxe, sexa aleatoria, deliberada ou teórica–, así como unha lóxica subxacente á distribución cronolóxica dos episodios de compilación de datos. Isto obriga a pensar coidadosamente que perspectiva (ou perspectivas) sobre o que se está a investigar pode resultar máis reveladora.

- O uso da **triangulación**. A triangulación consiste en combinar diversos tipos de datos –ou, por veces, diversos xeitos de observalos– para responder as cuestións que a investigación presenta. Denzin (1989) describe catro tipos de triangulación: a triangulación metodolóxica, que combina diferentes métodos de investigación; a triangulación de datos, que combina datos de máis dunha fonte; a triangulación de investigadores, que fai que máis dun investigador observe os datos para replicar ás interpretacións doutros investigadores; e a triangulación teórica, que contempla os datos desde diversas posturas teóricas para ver en que medida resultan adecuados e mais para comprender o xeito en que a consideración dos datos desde diferentes presupostos pode afectar ao xeito en que se comprenden. Destes catro tipos, os dous primeiros son os máis usados nas avaliacións gobernamentais.

- A atención á **validez** interna e externa dos resultados. En *Experimental and Quasi-Experimental Designs for Research* (1966), Campbell e Stanley chamaron *validez interna* á medida en que un experimento se controla adecuadamente e *validez externa* á medida en que son amplamente aplicables os resultados dun experimento. Por exemplo, nos plans de avaliación, é a miúdo imprescindible prestar atención ao 'peso morto' e ás externalidades para obter resultados válidos.

- A atención á **fiabilidade** dos resultados, de maneira que o instrumento de medición utilizado na avaliación, for cualitativo ou cuantitativo, dea resultados coherentes, estables e uniformes durante sucesivas observacións ou medicións efectuadas en idénticas condicións.

- A atención ao principio de **proporcionalidade**. Simplemente apunta á necesidade de que o traballo de avaliación corresponda á escala da intervención.

- A atención á **obxectividade** e á **integridade**. Os individuos que leven a cabo o traballo de avaliación deben carecer de impedimentos que empezan a súa obxectividade e actuar con integridade nas súas relacións con todas as persoas afectadas.

- A produción de resultados **oportunos**, **relevantes** e **verosímiles**. Os resultados da avaliación deben satisfacer as necesidades do organismo que a encargou e facerse públicos no momento máis oportuno para contribuír á toma de decisións administrativas. A evidencia debe ser suficiente en relación co contexto da toma de decisións; os resultados deben ser relevantes para os temas que se tratan e deducirse da evidencia. Os resultados da avaliación deben ser manifestamente útiles para os xestores á hora de mellorar o rendemento e de informar sobre os resultados obtidos.

- O **estilo accesible**. Os informes deben ser concisos e estar claramente redactados; non deben incluír máis información ca a necesaria para unha comprensión adecuada dos resultados, das conclusións e das recomendacións; deben presentar as conclusións e as recomendacións de maneira que se deduzan loxicamente dos resultados da avaliación; e deben tamén expoñer claramente os límites da avaliación en canto ao alcance, aos métodos e ás conclusións.

Notas

1 Para unha discusión máis detallada deste tema, véxase a última sección deste artigo.

2 Aínda que as avaliacións do impacto deberían cubrir os resultados duros e os brandos (máis intanxibles), a bibliografía respecto diso tende a centrarse nos primeiros: unha bibliografía recente recompilada por Lloyd e O'Sullivan (2004) revelaba que, de feito, hai moi poucas referencias á medición dos resultados brandos ou 'distancia percorrida'. Este feito contrasta cos diversos modelos prácticos de medición da distancia percorrida que se describen no seu traballo, o cal indica que a bibliografía académica e de investigación de políticas sobre este tema aínda non está ao mesmo nivel ca a práctica actual na administración de proxectos.

3 Esta tarefa non está exenta de problemas: por exemplo, hai que considerar coidadosamente que programa 'alternativo' se lle ofrecerá ao grupo de comparación (ou control), dado que isto definirá a hipótese. Idealmente, o grupo de control continuaría como o faría se non existise o programa. Con todo, hai motivos polos que isto pode non suceder:

1. O grupo de control pode decatarse da existencia da intervención, o cal pode afectar ao seu comportamento; por exemplo, os seus membros poden pospoñer as súas actividades de procura de emprego ata reunir as condicións para ser beneficiarios do novo programa.

2. Ao grupo de control pode ofrecérselle os 'mellores' servizos dispoñibles na actualidade como alternativa á iniciativa da política. Se, en ausencia do novo programa, non existen procedementos para informar a xente destes servizos, o grupo de control non será comparable coa actual situación.

3. A intervención pode ter efectos que afecten indirectamente ao grupo de control; por exemplo, cambiando a actitude dos empregadores locais cara aos ex-delincuentes cando se leve á práctica un programa de emprego para este grupo.

Outros problemas relacionados son o 'nesgo na posta en práctica' e o 'nesgo na cola'. O nesgo na posta en práctica dáse se, no contexto dun ensaio, non se pode poñer en práctica un programa como se poría nun contexto máis natural. O nesgo na cola pode darse porque nun ensaio só unha porcentaxe da poboación se ve afectado pola nova intervención, e isto pode darlle unha vantaxe inxusta con respecto ao grupo de control, cousa que non ocorrería se o programa se levase á práctica por completo.

Estes nesgos potenciais son moi difíciles, se non imposibles, de cuantificar. Os avaliadores teñen que esforzarse en minimizar os nesgos onde sexa posible, pero idealmente tamén terían que ser capaces de emitir xuízos informados acerca da medida en que poden ser problemáticos os nesgos.

4 As nosas descricións baséanse en gran medida en Purdon (2002).

5 En gran parte da nosa descrición seguimos a Davies (2004b).

6 Véxase, por exemplo, a afirmación de Subirats (1994: 9) segundo a cal "cando se clasifica unha política como éxito ou fracaso, isto a miúdo indica que se considerou desde unha estreiteza de miras centrada na xestión, máis preocupada por cumprir os fins internos da política ou por exercer un control administrativo efectivo ca pola capacidade do programa para responder ás necesidades dos diversos individuos e grupos afectados". Isto pasa por alto que as avaliacións de calidade –estean ou non dentro do paradigma tradicional– tamén deben valorar a relevancia das diversas iniciativas, así como a eficacia e mais a eficiencia. Esa valoración da relevancia é na actualidade un requisito para, por exemplo, todas as avaliacións que se acometen na Comisión Europea.

REFERENCIAS BIBLIOGRÁFICAS

Adler, M. e Ziglio, E. 1996. *Gazing into the Oracle: The Delphi Method and its Application to Social Policy and Public Health*. London: Jessica Kingsley Publishers.

Alkin, M. C., Hofstetter, C. H. e Ai, X. 1998. 'Stakeholder Concepts in Program Evaluation', en A. Reynolds e H. Walberg. *Advances in Educational Productivity,* vol. 7. Greenwich: JAI Press.

Blundell, R. e Costa Dias, M. 2000. 'Evaluation Methods for Non-Experimental Data', *Fiscal Studies,* 21 (4).

Boruch, R. F. 1997. *Randomized Experiments for Planning and Evaluation: A Practical Guide.* Newbury Park: Sage.

Bryson, A., Dorsett, R. e Purdon, S. 2002. *The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies.* Working Papers, 4. London: UK Department for Work and Pensions.

Campbell, D. T. e Stanley, J. C. 1966. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally College Publishing Company.

Cantrill, J. A., Sibbald, B. e Buetow, S. 1996. 'The Delphi and Nominal Group Techniques in Health Services Research', *The International Journal of Pharmacy Practice,* 4.

Chen, H. T. 1990. *Theory Driven Evaluations.* Thousand Oaks: Sage Publications.

Chen, H. T. e Rossi, P. H. 1983. 'Evaluating with Sense: A Theory-Driven Approach', *Evaluation Review,* 7 (3).

Connell, J. P. e Kubisch, A. C. 1995. 'Applying a Theory of Change Approach to the Evaluation of Comprehensive Community Initiatives: Progress, Prospects and Problems', en Fulbright-Anderson, K., Kubisch, A. C. e Connell, J. P. (eds.). *New Approaches to Evaluating Community Initiatives.* Washington: The Aspen Institute.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Light, R. J., Louis, T. A. e Mosteller, F. 1992. *Meta-Analysis for Explanation.* New York: Russell Sage.

Coote, A. e Lenaghan, J. 1997. *Citizens' Juries: Theory into Practice.* London: Institute for Public Policy Research.

Critcher, C. e Gladstone, B. 1998. 'Utilizing the Delphi Technique in Policy Discussion: A Case Study of a Privatized Utility in Britain', *Public Administration,* 76.

Davies, S., Elizabeth, S., Hanley, B., New, B. e Sang, B. 1998. *Ordinary Wisdom: Reflections on an Experiment in Citizenship and Health.* London: Kings Fund.

Davies, P. T. 2004a. 'Is Evidence-Based Government Possible?' Jerry Le Lecture, pronunciada no 4º Annual Campbell Collaboration Colloquium, Washington, 18-20 de febreiro, 2004.

Davies, P. T. 2004b. *The Magenta Book. Guidance Notes for Policy Evaluation and Analysis.* London: UK Government Chief Social Researcher's Office.

Deeks, J. J., Altman, D. G. e Bradburn, M. J. 2001 'Statistical Methods for Examining Heterogeneity and Combining Results from Several Studies in Meta-A nalysis', en Egger, M., Davey Smith, G. e Altman, D. G. (eds.). *Systematic Reviews in Health Care: Meta-Analysis in Context.* London: BMJ Publishing Group.

Denzin, N. K. 1989. *The Research Act: A Theoretical Introduction to Sociological Methods.* Englewood Cliffs: Prentice Hall.

Eisner, E. W. 1994 *The Educational Imagination: On the Design and Evaluation of School Programs.* New York, Toronto: Macmillan.

European Commission 1999. *Indicators for Monitoring and Evaluation: an Indicative Methodology.* The New Programming Period 2000-2006, Methodological Papers. Working Paper, 3. Brussels: DG XVI, Regional Policy and Cohesion, European Commission.

European Commission 2000. *The Mid-Term Evaluation of Structural Funds Interventions.* The New Programming Period 2000-2006, Methodological Papers. Working Paper, 8. Brussels: DG XVI, Rexional Policy and Cohesion, European Commission.

Fetterman, D. S. 1998. *Ethnography: Step by Step.* Thousand Oaks: Sage.

Fetterman, D. M., Kaftarian, S. J. e Wandersman, A. 1996. *Empowerment Evaluation: Knowledge and Tools for Self-Assessment and Accountability.* Thousand Oaks: Sage.

Fishkin, J. 1995. *The Voice of the People.* Yale: Yale University Press.

Funnel, S. 1997. 'Program Logic: An Adaptable Tool for Designing and Evaluating Programs', *Evaluation News and Comment,* 6 (1).

Gibson, T. 1998. *The Do-ers' Guide to Planning for Real.* Neighbourhood Initiatives Foundation.

Greenberg, D. H. e Morris, S. 2003. *Large Scale Social Experimentation in Britain: What Can and Cannot Be Learnt from the Employment Retention and Advancement Demonstration.* Occasional Papers, 3. London: UK Government Chief Social Researcher's Office.

Greene, J. C., Benjamin. L. e Goodyear, L. 2001. 'The Merits of Mixing Methods in Evaluation', *Evaluation,* 7 (1).

Guba, E. G. e Lincoln, E. S. 1989. *Fourth Generation Evaluation.* Newbury Park: Sage.

Hammersley, M. e Atkinson, P. 1995. *Ethnography: Principles in Practice.* London: Routledge.

Heckman, J. 1995 *Instrumental Variables: A Cautionary Talle.* Technical Working Paper, 185. Cambridge: NBER.

Heckman, J. J. e Smith, J. A. 1999. *The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies.* Working Papers, 6983. US National Bureau of Economic Research, Inc.

Lloyd, R. e O'Sullivan, F. 2004. *Measuring Soft Outcomes and Distance Travelled: A Practical Guide.* London: UK Department for Work and Pensions.

Majone, G. D. 1989. *Evidence, Argument and Persuasion.* New Haven: Yale University Press.

Marradi, A. 1990. 'Classification, Typology, Taxonomy', *Quantity and Quality,* XX (2).

Morton, S. 1999. *Systematic Reviews and Meta-Analysis.* Workshop Materials on Evidence-Based Health Care. San Diego, La Jolla: University of California.

Nagel, S. (ed.) 1990. *Policy Theory and Policy Evaluation: Concepts, Knowledge, Causes and Norms.* New York: Greenwood.

Owen, J. M. e Rogers, P. J. 1999. *Program Evaluation, Forms and Approaches.* London: Sage.

Park, A., Jowell, R. e McPherson, S. 1999. *The Future of the National Health Service: Results from a Deliberative Poll.* London: Kings Fund.

Patton, M. Q. 2002. *Qualitative Research & Evaluation Methods.* London: Sage

Pawson, R. e Tilley, N. 1997. *Realistic Evaluation.* London: Sage.

Purdon, S. 2002. *Estimating the Impact of Labour Market Programmes.* Working Paper, 3. London: UK Department for Work and Pensions.

Rodgers, P. J., Petrosino, A., Huebner, T. A. e Hacsi, T. A. 2000. *New Directions for Evaluation: Program Theory in Evaluation-Challenges and Opportunities.* San Francisco: Jossey Bass Publishers.

Rosenbaum, P. R. e Rubin, D. B. 1983. 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika*, 70.

Rossi, P. H., Freeman, H. E. e Lipsey, M. W. 1999. *Evaluation: A Systematic Approach.* Newberry Park: Sage Publications.

Scriven, M. 1972. 'Proles and Cons about Goal-Free Evaluation', *Evaluation Comment,* 3.

Seargeant, J. e Steele, J. 1998. *Consulting the Public: Guidelines and Good Practice.* London: Policy Studies Institute.

Slavin, R. E. 1984. 'Meta-Analysis in Education: how has it been used?', *Educational Researcher,* 13.

Slavin, R. E. 1986. 'Best Evidence Synthesis: An Alternative to Meta-Analysis and Traditional Reviews, *Educational Researcher,* 15.

Subirats, J. 1994. *Policy Instruments, Public Deliberation and Evaluation Processes.* Estudo-Working Paper, 1994-51. Madrid: Instituto Juan March de Estudios Avanzados en Ciencias Sociales.

Treasury Board of Canada Secretariat 2001. 'Evaluation Policy'. Dispoñible en http://www.tbs-sct.gc.ca/pubs_pol/dcgpubs/TBM_161/ep-pe1_e.asp#_Toc505657347http:www.tbs-sct.gc.capubs_poldcgpubsTBM_161ep-pe1_e.asp#_Toc505657347.

Weiss, C. H. 1997. 'Theory-Based Evaluation: Past, Present and Future', *New Directions for Evaluation,* 76.

White, H. 2005. *Challenges in Evaluating Development Effectiveness.* IDS Working Papers, 242. Brighton: University of Sussex.

White, C., Elam, G. e Lewis, J. 1999. *Citizens' Juries: an Appraisal of their Role.* London: Cabinet Office.

World Bank 2004. *Monitoring and Evaluation: Some Tools, Methods and Approaches.* Washington: World Bank Evaluation Operations Department.