Avaliação dos programas de saúde: perspectivas teórico metodológicas e políticas institucionais

Evaluation of health programs: theoretical-methodological prospects and institutional policies

Zulmira Maria de Araújo Hartz [1]

Abstract *This paper presents the theoretical-methodological perspective for health evaluation taking into account the political-decision making context of a possible governmental organization by programs and a management system by outcomes in the next five years. The main issues discussed herein are arranged into three groups: 1) the need for a logical or theoretical model guiding the evalution process; 2) the requirement for methodological plurality, given the context-oriented programmatic actions; and 3) the complexity of outcome measures; and the mandatory nature of institutional devices that regulate evaluations studies, thus guaranteeing the quality and utility of the final output. At the end, this paper emphasizes the importance of an evaluation research being fostered for the construction of scientific evidences in public health and its applications for the formation of evaluators.*

Key words *Health Evaluation; Health Program Evalution; Outcome-based Evaluation*

Resumo *Este artigo apresenta as perspectivas teórico-metodológicas da avaliação em saúde considerando o contexto político-decisório de uma possível estruturação governamental por programas e uma gestão por resultados no próximo qüinqüênio. As principais questões discutidas se organizam em três eixos: 1) a necessidade de um modelo teórico orientando o processo de avaliação; 2) a exigência de pluralidade metodológica dada à contextualização das ações programáticas a complexidade das medidas de resultados; e 3) a obrigatoriedade de dispositivos institucionais que regulamentem os estudos de avaliação garantindo a qualidade e utilidade do produto final. Ao final do texto se focaliza a importância de se fomentar uma pesquisa avaliativa voltada para a construção de evidências científicas em saúde pública e suas implicações para a formação de avaliadores.*

Palavras-chave *Avaliação em Saúde; Avaliação dos Programas de Saúde; Avaliação baseada em Resultados*

[1] Departamento de Epidemiologia e Métodos Quantitativos em Saúde, Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz, Rua Leopoldo Bulhões, 1480, 8º andar, 31041-210 Manguinhos Rio de Janeiro, RJ, Brasil zulmira@ensp.fiocruz.br

Introdução

In case you haven't noticed, things have changed significantly in the world of service delivery and program evaluation over last few years (Schalock, 1995 p. 3-4).

Considerado por Alves (1998) como uma *tentativa revolucionária de ordenar o gasto público* o Decreto Presidencial[1], referente à reestruturação federal por programas a partir de 2000, inspirada nos resultados da reforma administrativa dos Estados Unidos conduzida por Al Gore, não deve permanecer a "novidade invisível", com que se preocupa o autor, particularmente para nós profissionais de saúde pública que temos acumulado experiência no planejamento, gestão e avaliação das ações programáticas incorporadas em nosso modelo assistencial. A legislação americana aprovada em 1993 pelo Congresso (*Governement Performance and Results Act* – GPRA), tem similaridade com as propostas de reforma de diversos países desenvolvidos como a Grã-Bretanha, Canadá, Austrália e França que tentam imprimir na administração pública uma gestão por resultados (*outcomes*) (Wholey, 1997; Osborne & Plastrik, 1998; Pollit, 1998; Wegener, 1998). Na literatura a legislação americana se traduz pelo incremento da produção tecno-científica das medidas de *performance* englobando os esforços empreendidos pelos governos para avaliar os efeitos de suas intervenções (Newcomer, 1997; Wegener, 1998). Entre as inúmeras vantagens que têm sido apontadas, como justificativa do novo modelo, se inclui o conhecimento do nível de resultados alcançados, através de múltiplas avaliações, que se constitui em um barômetro de progresso, motivador do *staff* e orientador de atividades futuras (Plantz et al., 1997).

Embora a reengenharia governamental promovida pelo GPRA não esteja completamente implantada antes de 2002, uma análise preliminar mostrou, como ponto positivo, a maior focalização no consumidor, mas alertou para o baixo nível de envolvimento federal com a avaliação, comparado ao requerido pelo plano (Wargo, 1995). Estes achados sugerem problemas com a mensuração de desempenho nas pesquisas avaliativas apontados como desafios de natureza política ou metodológica a serem desmistificados (Jorjani, 1998). Os governos locais da *Organization for Economic Cooperation and Development* (OECD) também vêm tendo dificuldades, pois embora exista alguma experiência em avaliação setorial, a gestão por resultados força uma ampliação do conhecimento sobre todos os serviços fornecidos (Wegener, 1998). Toulemonde et al. (1998) reiteram as dificuldades inerentes a esse tipo de avaliação em que as parcerias, altamente desejáveis para o sucesso do novo modelo, legitimamente diferem em seus pontos de vista complicando cada estágio do processo.

O objetivo deste texto é o de levantar problemas conceituais e operacionais que emergem dos relatos de avaliadores envolvidos nessas experiências, bem como as alternativas de equacionamento que nos podem servir de aprendizado, destacando exemplos concernentes ao setor de saúde. Os programas são compreendidos como o conjunto de ações visando a favorecer comportamentos adaptativos requeridos pelas diferentes áreas ou atividades humanas relacionadas com vida comunitária, escola, trabalho, saúde e bem-estar. Sua avaliação demanda procedimentos de investigação para a coleta sistemática de informação voltada para a tomada de decisão e melhoria das intervenções (Schalock, 1995).

A premissa de base do artigo é focalizar a avaliação de programas públicos, dirigida para a aferição de resultados, o que exige a construção de um modelo teórico, explicitando como se espera que o programa exerça sua influência, uma pluralidade metodológica, contemplando a contextualização organizacional e a existência de dispositivos institucionais que regulamentem o processo de avaliação, garantindo qualidade e utilidade do produto final.

Modelo teórico

...o papel primordial numa teoria são as relações entre os objetos ...e para falar em relação há que existir uma trama, uma teia, uma rede... para então se construir ou descrever o bordado. O bordado não é obrigado a ocupar todo o tecido, pode ser um percurso, pode ser pontuações (Valente, 1998 p. 9-10).

Tradicionalmente a pesquisa avaliativa já se referia à análise dos resultados ou efeitos "líquidos" dos programas, sendo criticada pela sua limitação em não considerar as modalidades de implantação nos diferentes contextos nem os mecanismos intervenientes associados a esses efeitos (*black-box experiment)*. Denis e Champagne (1997) fazem uma ampla revisão das críticas feitas a este tipo de avalia-

Ciência & Saúde Coletiva, 4(2):341-353, 1999

ção, onde uma intervenção é tratada simplesmente como variável dicotômica, impermeável aos meios em que ela é introduzida. Eles apresentam as vantagens de uma avaliação orientada por um modelo teórico (*theory-driven evaluation* – TDE) explicitando como o programa supostamente funciona. Reynolds (1998) contrasta esta abordagem com uma orientação pelo método onde a incerteza é reduzida apenas através da alocação aleatória dos grupos nos desenhos de estudos experimentais ou com o controle estatístico exercido no momento da análise dos dados. Se estes mecanismos têm importância inquestionável para reduzir viéses que comprometam a validade interna, não podem esclarecer o como e o porquê dos programas atuarem nos grupos populacionais. A possibilidade de dar estas respostas, principal vantagem da TDE, é uma contribuição fundamental para a reprodutibilidade ou validade externa das intervenções governamentais em larga escala.

A construção do modelo teórico incluiria as seguintes especificações: 1) o problema ou comportamento visado pelo programa, a população alvo e as condições do contexto; 2) o conteúdo do programa ou atributos necessários e suficientes para produzirem isolada e/ou integradamente os efeitos esperados (ingredientes ativos). Esta construção lógica pode derivar de várias fontes como os resultados de pesquisas prévias, teorias das ciências sociais, experiência dos gestores e avaliadores (Reynolds, 1998). O programa é então tratado em sua pluralidade e também na singularidade de seus subprogramas ou projetos, nunca se tendo apenas um efeito desejado para os participantes mas um conjunto de efeitos lógica e hierarquicamente articulados em uma série de "se-então" relações associando recursos, atividades produzidas e resultados de curto e longo prazo (Plantz et al., 1997). Mercier (1990) discute o potencial dessa avaliação de programas, ao verificar elementos de realidade em seu "meio natural", para o desenvolvimento ou reconstrução das próprias teorias. Para Scriven (1998) existiriam duas teorias: a interna, que explica como se processam os produtos (*outputs*), e a externa que explica como os produtos se tornam os almejados resultados. Trochim (1998) julga que para os cientistas sociais o conceito de teoria interna é equivalente à noção de *construct* do programa (vinculada à validade externa) e que a teoria externa corresponderia à noção de causalidade (validade interna do estudo) não vendo vantagens em se criar novas denominações.

Os modelos teóricos, "popularmente" conhecidos no Canadá como modelos lógicos, constituem uma exigência governamental para avaliação das intervenções federais, desde o início dos anos 80, e são considerados extremamente práticos pelos avaliadores, ajudando-os a estabelecer e testar a razão do programa bem como a conceber um instrumento de avaliação adequado (Montague, 1997). Na avaliação do programa canadense para prevenção do câncer do seio, Mercer e Goel (1994) reforçam o desenvolvimento do modelo lógico como uma etapa "crucial" na apreciação da avaliabilidade de qualquer programa. Os autores relatam uma demanda de avaliação dos resultados de um programa de *screening*, efetuado em mulheres da província de Ontário, cuja modelagem revelou a complexidade da intervenção que na realidade incluía outros cinco subprogramas ou atividades interdependentes (promoção da saúde, educação profissional, recrutamento e aderência, seguimento de casos, parcerias de avaliação e pesquisa), requerendo diversas abordagens para avaliá-lo. Hartz et al. (1997), ao avaliarem o programa materno-infantil no Nordeste do Brasil, também construiram um modelo teórico evidenciando as articulações existentes entre os vários subprogramas (pré-natal, imunização, controle das infecções respiratórias agudas das doenças diarreicas, etc.) e os contextos intra/inter-organizacionais, que orientaram a seleção dos indicadores e os níveis de análise necessários. Para Hennessy (1995) a relevância de um modelo teórico é de tal ordem que os avaliadores só deveriam avaliar políticas e programas que tenham explicitado sua teoria e as medidas ou indicadores correspondentes.

Pluralidade metodológica

O desenvolvimento da avaliação enquanto profissão depende criticamente de apresentar argumentos racional e empiricamente fundamentados sobre os métodos e estratégias a serem escolhidos em que situações e porquê (Shadish, 1997).

À medida que a atenção à saúde exige respostas às necessidades de populações específicas com maior vulnerabilidade ou alto risco, a avaliação de programas baseada em princípios epidemiológicos, necessários para determinar estratégias de maior efetividade é con-

sensualmente tida como indispensável. A epidemiologia é reconhecida como a disciplina capaz de identificar variações nas probabilidades de resultados benéficos ou adversos (Clement et al., 1995). No entanto, avaliar um programa é muito mais que apenas estimar os diferenciais de risco ao final de uma intervenção. As estratégias de avaliação e pesquisa devem ser uma expressão prática do quadro teórico construído e os estudos epidemiológicos nem sempre se adaptam à lógica dos programas. Breart e Bouyer (1991) ao analisarem os métodos epidemiológicos em avaliação mostram que, do ponto de vista prático, várias circunstâncias obrigam a saída do modelo ideal dos ensaios aleatórios quando se está no campo da avaliação. Citam como exemplos a necessidade de respostas urgentes para tomada de decisão; a promulgação de um regulamento ou lei de amplo alcance, sem possibilidade de documentar a situação precedente; ou a limitação ética de se restringir o uso de uma tecnologia, com suficiente evidência de maior resolutividade, para se constituir um grupo controle de não expostos. Prost (1997) ao enumerar as diversas vantagens do uso da epidemiologia para o planejamento e tomada de decisão nas políticas sanitárias, por considerá-la um instrumento da quantificação dos fenômenos do grupo no que concerne à saúde, alerta que a "matematização triunfante" pode excluir da análise tudo que não é diretamente quantificado tornando de difícil tradução o "rigor gelado" dos números em conceitos ou argumentos inteligíveis pois uma decisão se baseia também em julgamentos de valores individuais e sociais.

Potvin et al. (1994) descartam a possibilidade de se restringir avaliações de programas comunitários de saúde, que têm como pressuposto a participação da população em todos os níveis e o estímulo a inovações contextuais, ao "credo epidemiológico ou quase-experimental" que tradicionalmente orienta a pesquisa avaliativa. Eles exemplificam com um modelo alternativo aplicado ao programa canadense de redução dos fatores de risco das doenças cárdio-vasculares usando um instrumento cuja construção foi localmente negociada. Em outro artigo Potvin (1990) já advertia que a abordagem experimental exigiria, entre outras condições, homogeneidade populacional, padronização e estabilidade dos meios de intervenção ao longo do tempo, a que um programa em saúde comunitária dificilmente se prestaria, dado seu caráter interativo (negociação entre parcerias) e iterativo (entre a ação e seus resultados).

A análise das intervenções sócio-sanitárias também colocaria pelo menos três dificuldades para os avaliadores de programas no seu "meio natural": 1) multiplicidade de objetivos adaptados à situação dos clientes; 2) reconhecimento de fronteiras das intervenções pelo aporte de recursos locais; e 3) duração variável de exposição ao programa, indo desde o atendimento episódico de uma crise ou problema agudo ao seguimento de longo prazo, com finalidades amplas como a autonomia e as melhorias da qualidade de vida (Mercier, 1990). A exigência de contextualização dos resultados observados é de tal ordem que, ao tema da última conferência da Sociedade Européia de Avaliação – *What Works for Whom,* que procurou reforçar a condicionalidade da efetividade, se propôs acrescentar a expressão "em que circunstâncias" como agenda não apenas da conferência mas de toda a avaliação (UKES, 1997).

Nessa perspectiva, os estudos de caso (*case-study research),* com múltiplos níveis de análise imbricados são fortemente recomendados dado que o objeto de investigação é de grande complexidade, a tal ponto que o fenômeno de interesse não se distingue facilmente das condições contextuais, necessitando informações de ambos. A replicação dos estudos de caso é paralela à idéia de um *cross-experiment* mas a generalização é feita pelo modelo teórico e não pelo processo amostral (Yin, 1993). A validade interna ou seja, a segurança com que se pode estabelecer uma ligação de causalidade entre o programa e seus resultados em um estudo de caso, depende de dois fatores: a qualidade e a complexidade da articulação teórica subjacente ao estudo e à adequação entre o modo de análise escolhido e o modelo teórico. É o grau de conformidade entre o conjunto de pressupostos do modelo e a realidade empírica observada que permitem fazer um julgamento sobre seu valor explicativo (Denis & Champagne, 1997).

Os estudos de caso têm sido ainda indicados para avaliar inovações programáticas com pouco conhecimento de sua eficácia, como a prioridade nacional da prevenção de drogas nos Estados Unidos (Yin, 1993) facilitando a construção de modelos lógicos, para intervenções subseqüentes, apoiados nas evidências acumuladas. Um dos pontos que tornaram os estudos de caso o modelo adequado nessa pes-

Ciência & Saúde Coletiva, 4(2):341-353, 1999

quisa foi o fato de não poder se limitar os jovens como unidade de análise sendo necessárias pelo menos quatro unidades: 1) os projetos, com diferentes finalidades e populações-alvo, como o da Flórida que pretendia reduzir disfunções sociais de mães adolescentes usuárias de drogas, com dupla intervenção/observação cobrindo a gestação e a primeira infância dos bebês; 2) as organizações (que operavam projetos) verificando hipotéticas condições técnico-estruturais e de relações comunitárias indispensáveis para este tipo de intervenção com financiamento público; 3) o programa propriamente dito (total de projetos financiados), como o do *High-Youth Program do U.S. Department of Health and Human Services,* que durante três anos aportou 55 milhões de dólares em 300 projetos de demonstração contratualizados e operados localmente; 4) o esforço nacional, reconhecendo-se que os fatores de risco são vinculados a condições de vida de minorias étnicas demandando uma análise que demonstre medidas políticas e sócio-econômicas capazes de agir sobre os determinantes comportamentais coerentes com as ações programáticas específicas.

Para Shea et al. (1995) sempre que um programa envolve diversas comunidades, o desenho multicêntrico em estudos de caso proposto por Yin (1989), é particularmente apropriado permitindo a descrição e explicação dos elementos de sucesso da implantação ao captar a intervariabilidade. As avaliações multicêntricas ou *multi-site evaluation* começam a aparecer então como uma das opções do campo metodológico quando os programas são implantados em várias localizações geográficas. As principais vantagens, como ocorre nos ensaios clínico-terapêuticos, se referem ao aumento do tamanho das amostras e ao poder de generalização. Pode-se agregar os dados populacionais, constituindo uma única unidade de análise, ou fazer uma agregação de médias e sua análise de variância intra/inter localidades. Sinacore e Turpin (1991) julgam que esses desenhos precisam de aprimoramento nas análises estatísticas e de adaptações para que sejam aplicados pelos avaliadores que devem responder também a uma variedade de demandas político-sociais a serem destacadas. A área da saúde é considerada a mais promissora para sua aplicação pela tradição já mencionada na pesquisa biomédica, disponibilidade de grandes bancos de dados, necessidade de maior padronização pelos sistemas multi-institucionais de assistência e a crescente evidência que seus efeitos variam através das regiões (Freedman, 1991).

Sobre as novas perspectivas no debate quanti-qualitativo dos últimos 20 anos, Gendron (1996) analisa os principais discursos que alimentaram o confronto entre os que julgavam impossível esta aliança, dada a incomensurabilidade entre elementos que não têm a mesma medida do ponto de vista paradigmático e o discurso da compatibilidade de outros autores considerando a "incomensurabilidade" exagerada e inexata. Apelando para a complexidade moriniana, propõe a noção de complementariedade das abordagens em que o pesquisador deve utilizá-la deixando a cada uma suas próprias potencialidades. Esta articulação sem fusão coincide com o referencial do modelo de avaliação com estudos de casos múltiplos e níveis de análise imbricados que temos trabalhado (Hartz et al., 1997). A expressão "monismo epistemológico", utilizada por Groulx (1997), é uma contestação ainda mais radical ao que chama de dualismo metodológico, tendo como um de seus pilares de sustentação o pensamento de Bourdieu que *recusa toda separação entre a epistemologia das ciências do homem e das ciências da natureza ...considerando a prática das ciências humanas e naturais com uma mesma racionalidade de provas que se atualiza na acumulação de conhecimentos e integração teórica dos pontos de vista existentes* (p. 49-50). Os argumentos de Groulx (1997) nos alertam para a falsa divisão ou concorrência de dualismos como a análise estatística e interpretativa, os inquéritos e as monografias, que, muitas vezes, serviriam para fortalecer determinados grupos ou abordagens no mercado das subvenções de pesquisa. Reichardt e Rallis (1994) concordam que uma possível cooperação entre essas diferentes tradições, parcialmente adversárias, enriqueceria o campo da pesquisa em geral e o da avaliação. A multiplicação dos estudos no último qüinqüênio mostra que este caminho, fortalecendo o aspecto conciliador, já vem sendo percorrido sem que se vislumbre o término do debate.

Riggin (1997) acredita que os avaliadores têm aprendido a combinar informação quantitativa e qualitativa o que é não só desejável como inevitável. A necessidade de se defender sua interdependência já foi substituída pela preocupação: como operacionalizá-la? Caracelly e Greene (1997) também concordam que o campo da avaliação tem evoluído do deba-

te paroquial para uma perspectiva mais "ecumênica" que considera o potencial de múltiplos questionamentos, com metodologias diversas, em função do propósito da indagação e da prática avaliativa. Nesta mesmo linha do "ecumenismo metodológico" existem expressões cujo significado merece ser examinado (Haldemann & Levy, 1996). Uma delas é a triangulação, concebida como sendo a comparação de pelo menos dois pontos de vista sobre a realidade estudada. A triangulação pode se fazer com os dados (diferentes eixos ou unidades temporais, espaciais e populacionais), pesquisadores, teorias ou métodos (desenhos de estudo integrados ou relacionados explicitamente a um mesmo objeto).

O fundamental é compreender que a multiplicação de pontos de vista não garante uma melhor validade. Não é a justaposição de instrumentos, mas sua integração pelo pesquisador em torno da lógica de um "referente comum" que constitui a prova da qualidade de uma pesquisa multimétodos. A pesquisa buscará obter uma cobertura mais extensa e aprofundada de um mesmo objeto pela maior variedade de informações. Os critérios para combinação são dados pelos conceitos e teorias que suportam a pesquisa (Haldemann & Levy, 1996). Um aspecto interessante é a noção de "pragmatismo teórico" evocada por Datta (1997a) em que o foco primário são os resultados a ser obtidos com os métodos escolhidos e não as bases epistemológicas dos quais emergiram. O termo "pragmático" corresponde a opções metodológicas que sejam práticas (operacionais), contextualizadas (levando em conta oportunidades e limites da situação em que emerge a demanda) e conseqüênciais no formato conceitual e padrões escolhidos. Chen (1997) descrevendo as contingências favoráveis à mixagem de métodos caracteriza-as como aquelas que requerem informação precisa e contextual, com disponibilidade parcial de dados confiáveis e com características de sistema aberto e fechado. O importante é que os dados coletados estejam teoricamente aninhados e relacionados no quadro lógico.

Exemplos como os de Datta (1997a) e de Chen (1997) ilustram este processo de articulação em função da modelagem teórica do projeto. O primeiro estudo, a avaliação de um dos diversos programas de sobrevivência infantil (Indonésia), financiados pela *Agency for International Development* na década de 80, se apresentava excepcionalmente desafiador. Dois desses desafios são bastante freqüentes: detectar mudanças em um programa que apenas apoiava ações já em funcionamento; e atribuir resultados porque outros doadores e/ou mudanças nacionais poderiam ter acelerado ou retardado o progresso observado. O tempo de permanência no local era de apenas três semanas com uma equipe de quatro pessoas. A Indonésia tinha 175 milhões de habitantes vivendo em 6.000 ilhas dispersas no oceano. Desenhos ideais como coorte ou séries temporais com a combinação dos componentes introduzidos eram absolutamente impossíveis. Com a utilização de diversas técnicas qualitativas (análise de documentos, entrevistas) e quantitativas (dados secundários de estudos prévios, da literatura e de outras fontes nacionais ou locais) foi possível concluir sobre a credibilidade das respostas às questões formuladas pois as informações de múltiplas fontes foram claramente buscadas e "tecidas" juntas (Datta, 1997a).

O outro estudo foi um programa para redução do abuso de drogas em 734 escolas de Taiwan em que, dada a falta de recursos, a amostra se restringiu a 31 escolas (estratificadas por tamanho e regime de estudo). A complementação do inquérito quantitativo entre os alunos, com entrevistas dos docentes, permitiu observar a coerência dos resultados demonstrando a robustez global do desenho apesar de se ter restringido a coleta de dados a um dia/escola visitada (Chen, 1997).

Completando este tópico observa-se que a possibilidade de se agregar diversos subestudos, cada um com seu próprio desenho, em uma única avaliação sob o formato de "caso", torna esta estratégia de pesquisa altamente promissora entre os cenários futuros (Yin, 1994). Datta (1997b) também considera a mixagem de métodos uma das tendências da avaliação no século XXI embora reconheça que atualmente a mixagem de dados (quanti-qualitativos) tem maior aceitação do que a de desenhos. Novamente os estudos de caso demonstram especial força articulando métodos e evidências quanti-qualitativas conjuntamente (Yin, 1994).

A mensuração de resultados

In the health care system this is the decade of outcomes... For population-based preventions to compete and attracting adequate funding, the

Ciência & Saúde Coletiva, 4(2):341-353, 1999

public health must move aggressively into this arena (Gold et al., 1997 p. 3)

Medir os efeitos atribuídos ao programa é o eixo das preocupações, dada à prioridade político-institucional e à complexidade das intervenções e de suas abordagens teórico-metodológicas. Wholey (1997) atribui o ressurgimento do interesse pelas medidas de *performance* nos Estados Unidos ao conjunto das reformas da administração pública introduzidas nos anos 60, culminando na qualidade total na década de 80 e com a gestão de resultados com o GRPA de 1993. Esta mesma tendência se verificou em outras "democracias industrializadas" como o Reino Unido e Austrália.

Para Scriven (1993) a avaliação de resultados usualmente produz *muito pouco muito tarde*. Entre outras justificativas cita: 1) a formulação de objetivos latentes diferentes dos manifestos, de caráter irrealistas ou doutrinários, nem sempre baseados em uma avaliação de necessidades; 2) a ausência da medida dos efeitos colaterais (independente dos resultados desejáveis terem sido atingidos) e do impacto sobre a população não usuária; 3) a falta de elementos de comparação, não sendo suficiente saber que o programa alcançou seus objetivos sem efeitos colaterais quando poderia ser feito melhor, com custo ligeiramente maior, similar ou até menor. Em síntese, a avaliação de programa não pode ser apenas alcance de objetivos precisando do monitoramento dos projetos correspondentes checados oportunamente na qualidade de seus processos e custos.

Aos problemas já levantados, acrescente-se o fato de que os resultados do programa no plano individual nem sempre correspondem em termos de comunidade sendo que algumas populações apresentam dificuldades específicas como usuários anônimos de serviços e demandante de intervenções emergenciais ou de curta duração (Newcomer, 1997). Entre as limitações das medidas de resultados encontram-se ainda os estudos de satisfação do usuário que tendem a superestimar efeitos de curto prazo, podem ser manipulados em inquéritos que induzem respostas desejadas pelos administradores (um grupo focal com boa representatividade para elaborar o instrumento minimizaria este viés) ou ignorar alguns grupos especiais (Wegener, 1998).

Nos estudos epidemiológicos, Bréart e Bouyer (1991) listam os indicadores mínimos necessários para medir os efeitos dos programas em concordância com as diretrizes anteriores: nível de risco das populações expostas e não expostas; aplicação da intervenção (cobertura, qualidade da atenção); efeitos positivos e negativos; outras intervenções que possam influenciar os resultados. É interessante o exemplo de Grémy et al. (1995) sobre potenciais efeitos negativos do programa de prevenção do câncer do seio na França, levando em conta o fato de que existem 200 mamógrafos na Inglaterra e 2.000 na França, para uma população apenas 25% maior e com taxas de riscos similares, temendo-se pela provável multiplicação de atos médicos desnecessários e pelo menor controle da qualidade com este volume de equipamentos.

De volta à discussão dos indicadores requeridos para avaliação de resultados, Schalock (1995) recomenda que sejam objetivos, mensuráveis e com os seguintes atributos: logicamente conectados ao programa; com abrangência multidimensional englobando o conjunto de preditores, como a percepção da qualidade de vida em saúde; valorização individual, na perspectiva de consumidor preocupado com a qualidade do serviço e não apenas de paciente ou cliente; observados longitudinalmente pois os efeitos, e conseqüentemente os custos a eles associados, variam em função do tempo.

Um bom exemplo do campo da saúde, que enfoca várias das questões aqui comentadas, é o relato de Roberts e Wasik (1996) sobre uma avaliação federal em 41 comunidades americanas com sistemas integrados de serviços de atenção materno-infantil. Os editais do programa se fizeram em duas etapas (1992-1993) totalizando o financiamento de 43 projetos. Embora os termos de referência contratualizados tivessem como proposta geral a "redução da mortalidade infantil e melhoria dos indicadores de saúde através da expansão e desenvolvimento de sistemas integrados de serviços dirigidos às famílias", ficou claro que as intervenções constituíam um conjunto diversificado de soluções comunitárias aos problemas, sem objetivos ou medidas comuns. Nenhum modelo teórico, que integrasse as diversas unidades de análise (criança, mães, famílias e comunidade) através das diferentes hierarquias administrativas (nacional, regionais e locais), foi proposto. Indicadores mais complexos de saúde como aferições do desenvolvimento infantil, do abuso e/ou negligência com crianças, estavam sempre ausentes. A variabilidade evidenciada dos projetos só per-

mitiu classificá-los enquanto estudos de caso individualizados, tornando impossível um estudo multicêntrico de análise da implantação. Os autores vêm como principal lição a necessidade que o nível federal tenha mecanismos para assegurar que o plano de avaliação esteja contemplado desde a fase conceitual dos projetos, com objetivos e metas comparáveis para avaliar o investimento nacional, e que se realize um esforço sistemático para desenvolver medidas de "construtos" associados ao conceito de sistemas integrados de serviços.

A experiência do Québec, ao definir as prioridades nacionais em saúde pública para o período 1997-2002, é outro exemplo mais recente em que ainda permanecem algumas das questões levantadas no estudo precedente. Considerada por Roy et al. (1998) como uma ocasião "imperdível" de se abrirem novas perspectivas na avaliação em saúde pública, os autores reconhecem que tem sido uma "odisséia" a gestão por resultados. Embora se inscreva em um quadro teórico já articulando, em uma cadeia de causalidade, as diversas prioridades com as estruturas e o processo das atividades produzidas e os resultados esperados no nível nacional, regional e local, não há consenso operacional entre as múltiplas parcerias, e os problemas da escolha dos indicadores persistem: medir bem as boas e mesmas coisas para assegurar a comparabilidade dos resultados. Um conceito original como "gestantes sub-escolarizadas em extrema pobreza" terá de ser convertido em "nascidos de mães com menos de 11 anos de escolaridade", em virtude da não disponibilidade de dados, o que se torna de um reducionismo questionável e pode comprometer a validade de construção de uma pesquisa que retrata a correspondência entre uma estratégia de medida e os conceitos a que se refere (Denis & Champagne, 1997). Na perspectiva adotada os resultados, sendo de importância crucial, devem ser claros, precisos e avaliáveis embora sejam também objetos de compromissos, negociações entre o desejável e o possível, traduzindo ideologias, crenças e relações de força. Todas estas dificuldades, no entanto, não devem impedir o processo de avaliação ainda que muitos dados sejam recolhidos com "entorses" aos critérios científicos puros (Roy et al., 1998, p.170). O papel dos indicadores, como facilitadores das mudanças organizacionais desejadas, capazes de convencer os que decidem e mobilizar a população para o alcance das prioridades nacionais do Québec, exige que tenham as seguintes características: provocar esperança, pois muitas vezes são mais desencorajadores que mobilizadores, os mercadores de problemas superando os de soluções; fazer convergir o maior número de atores sem necessidade do consenso de todos; reconhecer obstáculos e compreender resistências das parcerias; prever indicadores que testemunhem sucesso de curto prazo; privilegiar a comunicação interpessoal antes da mídia que raramente se interessa por boas novas (Lagarde, 1998).

Conclusões

O "estado da arte" sobre a avaliação de resultados em programas públicos mostra que o problema não é novo para os avaliadores mas o requisito de ubiqüidade no seu uso o é (Newcomer, 1997). O GPRA chegou mesmo a ser imaginado como "a lei do pleno emprego para os avaliadores de programa" (p. 1) devido aos desafios apresentados para os gestores em termos de sua medida. Esta massificação da pesquisa em avaliação com vistas à eficiência dos programas exige, em princípio, precaução e parcimônia por parte dos avaliadores (Hartz & Pouvourville, 1998). As conseqüências adversas dessa obrigatoriedade da avaliação começam também a preocupar a Nova Zelândia. O argumento de Bushnell (1998), justificando tais preocupações, é o de que se uma avaliação não alimenta decisões, não pode dar respostas claras (como alguns estudos de eficiência) ou já dispõe de inferências válidas que podem ser generalizadas, ela não deveria ser efetuada, pois avaliação por avaliação seria uma estéril utilização de recursos escassos. Aliás, um problema de maior relevância que raramente se enfrenta é o custo das medidas de *performance*. Quando o GPRA foi aprovado no congresso americano estimou-se que o mínimo necessário seriam 50 milhões de dólares anuais entre 1996-1998, e que se elevaria nos anos subseqüentes pois a informação de resultados requer estudos adicionais (Wholey, 1997).

Nos países em desenvolvimento, e não apenas no Brasil, esta direcionalidade para a avaliação de resultados é reforçada por organismos de ajuda internacional, como o Banco Mundial, principal provedor de assistência técnica e financeira nesses países, que passou a incluir a "capacidade em avaliação" entre as prioridades para a gestão de atividades do setor pú-

Ciência & Saúde Coletiva, 4(2):341-353, 1999

blico, visando a sustentabilidade dos programas (Piccioto, 1997), e vem diversificando as metodologias utilizadas para incorporar inquéritos voltados para o consumidor, com uso mais sistemático dos grupos focais e adaptação de técnicas de avaliação rápidas já utilizadas no campo da saúde (Anker et al., 1993). Falando da experiência britânica, também dirigida para o financiamento de projetos de desenvolvimento internacional, Foulkes (1998) considera que o "novo" é a detalhada atenção que se está dando ao investigar os progressos obtidos pelos programas a partir de um modelo lógico definindo antecipadamente não só o que se vai fazer mas como e quando se atingem os objetivos, e ressalta que julgamentos qualitativos contam tanto quanto a estatística, o desafio não estando em agregar números mas em fazer um julgamento baseado em evidências. O interesse em construir maior capacidade em avaliação nas estruturas administrativas dos países em desenvolvimento decorre da constatação de que só recentemente este tema começa a interessá-los (Khan, 1998). O autor coloca como pré-requisito de sucesso uma maior conscientização dos benefícios da avaliação e sua incorporação nas estruturas governamentais, com envolvimento de parcerias não governamentais, integrando o processo da reforma e não como uma atividade isolada.

Os manuais ou guias de avaliação de organismos internacionais (OECD, Banco Mundial, Comissão Européia) e dos países com maior grau de avanço neste processo mostram a necessidade de se ter disposições institucionais, enquanto mecanismos de regulação, indispensáveis para se avaliar programas públicos que orientem a tomada de decisões (avaliação interna ou externa, comitês de pilotagem, perfil de avaliadores, escolha de metodologias, qualidade dos estudos realizados). A análise desses textos efetuada por Perret (1998) mostra que a maioria dá preferência à avaliação externa, sem excluir a interna, como é o caso da Comissão Européia que vê na avaliação interna um instrumento de aprendizagem onde os gestionários são estreitamente associados ao porquê e ao como de suas atividades. Considera-se também que a auto-avaliação e a avaliação independente (externa) são distintas mas complementares (Banco Mundial). A França e a Itália garantem um caráter misto designando um comitê de pilotagem (ao contrário da América do Norte que recorre a avaliadores profissionais) com um responsável operacional. Alguns manuais recomendam a associação de avaliadores multidisciplinares para viabilizar abordagens metodológicas complexas sob a forma de equipes, consórcio, etc. Duffy (1994) comentando o uso da avaliação interna destaca como vantagens um maior custo-efetividade, dado seu posicionamento no *staff* organizacional, com possibilidade de melhor utilização de resultados inclusive de propôr estratégias mais adequadas, propiciando adesão a longo prazo. De outra parte, a necessidade de experiência em avaliação e um mínimo de credenciais acadêmicas bem como o fato de despertar suspeitas por abraçar idéias dos executivos ou receber pressões significativas, com maior freqüência de dilemas éticos, se constituem em desvantagens ou dificuldades a se ponderar. O que parece mais importante é alertar que existe sempre a dificuldade de se conciliar a independência do avaliador e o envolvimento nos serviços durante o processo de avaliação. Em qualquer circunstância, a qualidade da avaliação depende da competência dos avaliadores e da capacidade da instituição em considerar e utilizar apropriadamente os resultados (Corbeil & McQueen, 1991). Neste sentido os órgãos governamentais podem aproveitar o esforço feito pelas associações de avaliadores, que nos últimos anos vêm tentando estabelecer padrões de qualidade para a avaliação de programas, propondo critérios mínimos ou princípios com que os avaliadores possam orientar sua prática (CES, 1992) e institucionalizá-los.

Os Padrões de Avaliação de Programas[2] fornecem diretrizes e casos ilustrativos para auxiliar aqueles envolvidos em uma avaliação a alcançar cada um desses padrões. Os casos ilustrativos baseiam-se em uma infinidade de cenários educacionais, incluindo escolas, universidades, os campos da medicina e da saúde, as forças armadas, o comércio, a indústria, o governo e a polícia. Eles identificam princípios avaliativos com quatro atributos básicos que, quando empregados, poderiam resultar em uma melhor avaliação de programas: **utilidade:** a avaliação atende às necessidades de informação dos interessados; **praticabilidade:** a avaliação é realista, prudente, hábil e econômica (politicamente viável e eficiente, produzindo informações suficientemente úteis, de modo que os recursos gastos possam ser justificados); **propriedade:** a avaliação é conduzida de forma legal e ética, preocupando-se com todos aqueles envolvidos ou afetados por seus

resultados (as obrigações das partes devem ser acordadas por escrito, de modo que fiquem obrigadas a respeitá-las); **precisão:** a avaliação produz informações tecnicamente adequadas sobre os elementos que determinam o valor ou o mérito do programa. As conclusões são explicitamente justificadas, de modo que os usuários ou pares possam julgá-las em sua validade.

A importância de se considerar o conjunto de atributos como sinérgicos é fundamental pois validade não garante credibilidade e ambos não asseguram utilidade (Scriven, 1993).

Datta (1997b) propõe, para a verificação da qualidade dos estudos de caso, que além dos itens correspondentes às dimensões anteriores, se agregue a "eqüidade metodológica": garantir a mesma profundidade e transparência para a pluralidade de abordagens metodológicas evitando-se estudos qualitativos com embelezamento quantitativo ou vice-versa (p. 353).

Resta fazer uma discussão sobre os limites do "julgamento baseado em evidências" reclamado por Foulkes (1998). Existe consenso em uma gestão onde o que conta são os resultados, que precisa-se fazer mais e melhor uso de evidências e que os experimentos sejam complementados com outros desenhos empíricos e modelos teóricos que permitam dar transparência às intervenções como também estimar os efeitos de agregação (UKES, 1998). Nessa lógica a "pesquisa de síntese" para acumulação de evidências é atualmente uma das principais atividades da Divisão de Metodologias e Avaliação de Programas (PEMD), ligada ao *General Accounting Office* do Congresso Americano, considerada uma das organizações de mais alta qualidade no meio profissional (Cook, 1997). Um dos problemas desta abordagem é que a efetividade das ações programáticas tem como premissa a observação empírica dos resultados populacionais, teoricamente alcançáveis com base em "evidências científicas" prévias dos estudos experimentais de seus componentes, associados em uma suposta relação "causal", e que estas evidências são quase sempre incompletas ou insuficientes na saúde.

Se a contribuição do desenvolvimento da medicina-baseada em evidência com as iniciativas americanas como o *Office of Medical Applications Research* ou o apoio dado aos estudos que sintetizam conhecimentos produzidos como faz a *Agency for Health Care Policy and Research* se constituem avanços importantes para a pesquisa avaliativa (Novaes, 1996) ela é geralmente restrita ao campo biomédico. Uma saúde pública baseada em evidências ou *best practices* – avaliação sistemática dos achados de pesquisa orientando decisões para as práticas relacionadas a atenção a saúde – é ainda incipiente. Um dos maiores desafios é o da promoção da saúde, definida como o processo que favorece indivíduos e comunidades a melhorar a saúde aumentando seu controle sobre os diversos determinantes, o que torna a seleção das variáveis dependentes e independentes bem mais difícil e o tempo de espera dos resultados almejados bem mais longo do que a assistência médica e a prevenção de doenças (Jenicek, 1997). Este obstáculo não deve ser paralisante e a própria PEMD tem utilizado procedimentos de revisão sistemática adaptados da meta-análise, com pouco estudos quantitativos disponíveis, mas complementados com telefonemas a pesquisadores, consultores e busca de relatórios pouco divulgados (Cook, 1997). A *American Public Health Association* (APHA, 1990) faz as seguintes recomendações, em presença de evidências incompletas: 1) divulgar com os atores envolvidos as evidências disponíveis identificando-se os viéses dos estudos e suas implicações sobre os resultados; 2) considerar os benefícios da intervenção, métodos e custos de implantação prevendo-se revisão periódica das evidências disponíveis a serem incorporadas; 3) detalhar fatores que comprometem o sucesso do programa lembrando que a incerteza científica pode ser explorada por grupos de interesse disfarçando oposição; avaliar as estratégias de intervenção considerando os aspectos controversos da questão, se possível analisando diferentes áreas geográficas; 4) verificar se todos os requisitos éticos foram respeitados; 5) aumentar a compreensão dos usuários melhorando a comunicação científica; 6) fortalecer a formação das escolas de saúde pública na análise científica de evidências.

Concluindo esta sumarização de literatura esperamos ter deixado claro que, se a gerência unificada da organização governamental por programas facilita a cobrança de responsabilidades, pelos objetivos fixados, em "quem tem nome, telefone e endereço" (Alves, 1998), a atribuição dos resultados obtidos aos programas implantados exigirá múltiplos focos de avaliação, articulados por um modelo teórico ou lógico, de modo a se mostrar de forma coerente e convincente a presumida associação entre as intervenções e os estados de

Ciência & Saúde Coletiva, 4(2):341-353, 1999

saúde observados. A última recomendação da APHA aponta para a exigência de se atualizarem os conteúdos da formação de recursos humanos, privilegiando a análise de evidências em saúde pública, o que considero um ponto nodal mas insuficiente para implementar a avaliação dentro das perspectivas teórico-metodológicas aqui apresentadas. Concordo com Corbeil e McQueen (1991) quando dizem que a avaliação é um artesanato sutil combinando muitos ingredientes: análises, comunicações, expectativas, julgamentos, política e psicologia exigindo do avaliador integridade, flexibilidade, agilidade intelectual e criatividade. Parece, no entanto, que Chelimsky (1997), ao preconizar um treinamento em avaliação "politicamente realista", localiza a principal carência dos modelos de formação de avaliadores: omitir os determinantes políticos dos aspectos teórico-metodológicos fazendo crer que a uma boa avaliação se seguem decisões imediatas, desconhecendo que este é apenas um dos elementos (nem sempre o mais importante) da agenda política, demandando aos avaliadores perseverança e reforço contínuo de argumentos. Para a autora é também indispensável saber que as orientações ideológicas, privilegiando a implementação ou os cortes dos programas sociais, podem gerar quadros teórico-metodológicos similares para a pesquisa mas se baseiam em valores muitas vezes opostos (como é o caso dos democratas *versus* conservadores americanos) e têm implicações distintas para a vida pública. Nem a política nem a evidência que oferecemos para subsidiá-la é perfeita. Elas são iterativas e o desafio é compreender a força e vulnerabilidade de ambas para contribuir de forma significativa na formação e prática dos avaliadores a quem Chelimsky (1997) lembra: *ninguém está pedindo para sermos sublimes: somente sérios, com credibilidade e persistência* (p. 225).

Notas

1 "O PRESIDENTE DA REPÚBLICA...DECRETA:... Para a elaboração e execução do Plano Plurianual 2000-2003 e dos Orçamentos da União, a partir do exercício financeiro do ano 2000, toda ação finalística do Governo Federal deverá ser estruturada em programas orientados para a consecução dos objetivos estratégicos definidos para o período do plano" (Dec. 2.289 de 29/10/98).

2 Aprovado pelo *American National Standards Institute* (ANSI) como Padrão Nacional Norte-Americano (15/3/1994). *The Program Evaluation Standards* (1994), Thousand Oaks, C. A: Sage Publications

Referências

Alves MMA 1998. Novidade Invisível. *O Globo*, 1º de novembro, p. 4.

Anker M, Guidotti RJ, Orzeszyna S, Sapírie AS, Thuriaux MC 1993. Rapid evaluation methods (REM) of health services performance methodological observations. *WHO Bulletin OMS 71*(1):15-21.

APHA – American Public Health Association 1990. Public health policy – Making in the presence of incomplet evidence. *AJPH 80*(6): 746-750.

Breart G, Bouyer J 1991. Méthodes épidemiologiques en évaluation. *Révue d'Epidemiologie et Santé Publique 39*: S5-S14.

Bushnell P 1998. Does evaluation policies matter? *Evaluation 3*(4): 363-372.

Caracelly VJ, Greene JC 1997. Crafting mixed-method evaluation designs. *New Directions for Evaluation 74* Summer: 19-29.

CES – Canadian Evaluation Society 1992. Standards for program evaluation in Canada: a discussion paper. *The Canadian Journal of Program Evaluation 7*(1): 157-170.

Chelimsky E 1997. The political environment of evaluation and what it means for the development of the field, p. 53-68. In E Chelimsky & WR Shadish (orgs.) *Evaluation for the 21st Century*. Sage Publications, London.

Chen H-t 1997. Applying mixed methods under the framework of theory-driven evaluation. *New Directions for Program Evaluations 74*: 61-73.

Clement DG, Wan TTH, Stegall MH 1995. Evaluating health care programs and systems. An Epidemiologic Perspective, p. 79-99. In DM Oleske *Epidemiology and the Delivery of Health Care Services: Methods and Applications,* Plenum Press, New York.

Cook TD 1997. Lessons learned in evaluation over the past 25 years, p. 30-52. In E Chelimsky & WR Shadish *Evaluation for the 21st Century*. Sage, Thousand Oaks, CA.

Corbeil RC & Mcqueen C 1991. Improving the quality of evaluation, p. 196-213. In AJ Love *Evaluation Methods Sourcebook.* CES, Toronto.

Datta L-e 1997a. A pragmatic basis for mixed-method designs. *New Directions for Evaluation 74* Summer: 33-46.

Datta L-e 1997b.-Multimethod evaluations. Using case-studies together, p. 344-359. In E Chelimski & WR Shadish (orgs.) *Evaluation for the 21st Century*. Thousand Oaks, CA.

Denis JL, Champagne F 1997. Analise da implantação, p. 49-88. In ZMA Hartz *Avaliação em Saúde: dos Modelos Conceituais à Prática na Análise da Implantação de Programas.* Fiocruz, Rio de Janeiro.

Duffy BP 1994. Use and abuse of internal evaluation. *New Directions for Program Evaluation 64* Winter: 25-32.

Foulkes G 1998. Evaluation and international development: a britsh perspective. *Evaluation 3*(4): 359-362.

Freedman JA 1991. Multisite evaluations of health care policies and programs: new directions for program. *Evaluation 50* Summer: 97-108.

Gendron S 1996. L'alliance des approches qualitatives et quantitatives en promotion de la santé: vers uns complémentarité transformatrice. *Ruptures 3*(2): 158-172.

Gold MR, McCoy KL, Teutsch SM, Haddix AC 1997. Assessing outcomes in population health: moving the field forward. *American Journal of Preventive Medicine 13*(1): 3-5.

Gremy F, Manderscheid J-C, Penochet J-C 1995. Evaluation et qualité dans le domaine de la santé. *Santé Publique et Territoires, 10 ans de decentralisation.* ENSP, Rennes.

Groulx L-H 1997. Le débat qualitatif-quantitatif: un dualisme à proscrire? *Ruptures 4*(1): 46-58.

Haldemann V, Levy R 1996. Oecumenisme méthodologique et dialogue entre paradigmes. *Ruptures 3*(2): 244-255.

Hartz ZMA, De Pouvourville G 1998. Avaliação dos programas de saúde: a eficiência em questão. *Ciência & Saúde Coletiva III* (1): 68-82.

Hartz ZMA, Champagne F, Leal MC, Contandriopoulos A-P 1997. Avaliação do programa materno-infantil: análise de implantação em sistemas locais de saúde no nordeste do Brasil, p.19-28. In ZMA Hartz *Avaliação em Saúde: dos Modelos Conceituais à Prática na Análise da Implantação de Programas.* Fiocruz, Rio de Janeiro.

Hennessy M 1995. What works in program evaluation. *Evaluation Practice 16*(3): 275-278.

Jenicek M 1997. Epidemiology, evidence-based medicine and evidence based public health. *Journal of Epidemiology 7*(4): 187-197.

Jorjani H 1998. Demystifying results-based performance measurements. *Canadian Journal of Program Evaluation 13*(1): 61-95.

Khan MA 1998. Evaluation capacity building. An overview of current states, issues and options. *Evaluation 3*(4): 310-328.

Lagarde F 1998. Les indicateurs: source de statistiques ou source de changements? *Ruptures 5*(2): 173-178.

Mercer SL, Goel V 1994. Program evaluation in the absence of goals: a comprehensive approach to the evaluation-based breast screening program. *The Canadian Journal of Program Evaluation 9*: 97-112.

Mercier C 1990. Evaluation des programmes d'intervention en milieu naturel. *The Canadian Journal of Program Evaluation 5*(1): 1-16.

Montague S 1997. Les evaluateurs et la mesure du rendement: Mettre le modèle logique à disposition du gestionnaire. *Bulletin de la Societé Canadienne d'Évaluation 17*(2): 1-2.

Newcomer KE 1997. Using performance measurement to improve public and nonprofit programs. *New Directions for Evaluation 75* Fall: 1-14.

Novaes HMD 1996. Epidemiologia e avaliação em serviços de atenção médica: novas tendências na pesquisa. *Cadernos de Saúde Pública 12*(2): 7-12.

Osborne D, Plastrik P 1998. Repensando o serviço público. *Reforma Gerencial* março: 28-29.

Perret B 1998. Analyse comparéee de neuf guides d'évaluation, p.73-108. In CSE (Conseil Scientifique de l'Evaluation) *L'Évaluation en Developpement 1997*. La Documentation Française, Paris.

Picciotto R 1997. Evaluation in the Word Bank, p. 201-213. In E Chelimsky & WR Shadish (orgs.) *Evaluation for the 21 st Century*. Sage Publications, London.

Plantz MC, Greenway MT & Hendricks M 1997. Outcomes measurements: showing results in the nonprofit sectot. *New Directions for Evaluation 75* Fall: 15-30.

Pollit C 1998. Institutions et usages de l'évaluation au Royaume-Uni: vue d'ensemble, p. 26-33. In CSE (Conseil Scientifique de l'Evaluation) *L'Évaluation en Developpement 1997*. La Documentation Française, Paris.

Potvin N 1990. L'Évaluation de programme en santé communautaire: une question de negotiation. *The Canadian Journal of Program Evaluation 5*(1): 57-71.

Potvin L, Paradis G, Lessard R 1994. Le paradoxe de l'évaluation des programmes communautaires multiples de promotion de la santé. *Ruptures, revue transdisciplinaire en santé, 1*(1): 45-57.

Prost A 1997. La place de l'épidemiologie dans le processus de décision. *Cahiers Santé 7*: 61-64

Reichardt CS, Rallis SF 1994. The relationship between the qualitative and quantitative research traditions. *New Direction for Program Evaluation 61* Spring: 5-11.

Reynolds AJ 1998. Confirmatory program evaluation: a method for strengthening causal inference. *American Journal of Evaluation 19*(2): 203-221.

Riggin LJC 1997. Advances in mixed-method evaluation: a synthesis and comment. *New Direction for Evaluation 74* Summer: 87-94.

Roberts R, Wasik B 1996. Evaluating the 1992 and 1993 community integrated service system projects. *New Direction for Evaluation 69* Spring: 35-49.

Roy D, Fortin L, Potvin L, Valentini 1998. L'évaluation des priorités nationales dans un contexte de gestion par résultats. *Ruptures* 5(2):166-172.

Schalock RL 1995. *Outcome-Based Evaluation*. Plenum Press, New York, 242 pp.

Scriven M 1993. Hard-won lessons in program evaluation. *New Directions for Program Evalaution 58* Summer: 5-48.

Scriven M 1998. Minimalist theory: the least theory that practice requires. *American Journal of Evaluation 19*(1): 57-70.

Shadish WR 1997. Performance measurement and evaluation, 121-123. In E Chelimsky & WR Shadish (orgs.) *Evaluation for the 21 st Century*. Sage Publications, London.

Shea MP, Lewko JL, Flynn RA, Boschen KA, Volpe R 1995. Design and measurement considerations in evaluating integrated human service delivery systems. *Evaluation Practice 16*(3): 247-256.

Sinacore JM, Turpin RS 1991. Multiple sites in evaluation research: a survey of organizational and methodological issues. *New Directions for Program Evaluation 50* Summer: 1-18.

Toulemonde J, Fontaine C, Laudren E, Vincke P 1998. Evaluation in partnership. Practical suggestions for improving their quality. *Evaluation 4*(2): 171-188.

Trochim WMK 1998. An evaluation of Michel scriven's "the least theory that practice requires". *American Journal of Evaluation 19*(2): 243-249.

UKES – United Kingdom Evaluation Society 1997. *Newsletter 5* May.

UKES – United Kingdom Evaluation Society 1998. What works in public sector? *The Evaluator 2*: 6.

Valente JP 1998. *Sobre Modos de Transmissão da Matemática* (mimeo), 15 pp.

Wargo LJ 1995. The impact of federal government reinvention on federal evaluation activity. *Evaluation Practice 16*(3): 227-237.

Wegener A 1998. Evaluating competitively tendered contracts. local governments in comparative perspective. *Evaluation 4*(2): 189-203.

Wholey JS 1997. Trends in performance measurement. challenges for evaluators, p.124-133. In E Chelimski & WR Shadish (orgs.) *Evaluation for the 21st Century*. Sage, Thousand Oaks, CA.

Yin RK 1989. *Case Study Resarch: Design and Methods*. Sage, Newbury Park, CA.

Yin RK 1993. Case study designs for evaluating high-risk youth programs: the program dictates the design, p. 77-93. In *Applications of Case Study Research*. Sage, Newbury Park.

Yin RK 1994. Discovering the future of the case study method in evaluation research. *Evaluation Practice 15*: 283-290.