Psicologia: Teoria e Pesquisa
Out-Dez 2008, Vol. 24 n. 4, pp. 545-558

Avaliação de Programas Sociais: Comparando Estratégias de Análise de Dados

Luciana Mourão[1]
Universidade Salgado de Oliveira
Jacob A. Laros
Universidade de Brasília

RESUMO - Este artigo apresenta uma discussão sobre questões metodológicas em relação à avaliação de programas sociais, tomando como exemplo o caso do Plano Nacional de Qualificação do Trabalhador – PLANFOR. O delineamento da pesquisa foi de um estudo quase-experimental, com três grupos: dois experimentais e um controle. A pesquisa foi realizada durante os anos de 2001 a 2004, com coleta de dados antes e depois do programa, englobando 360 instituições, sendo uma por município, cobrindo todas as regiões brasileiras. Para a coleta de dados, utilizou-se um instrumento com indicadores quantitativos (sociais e econômicos). Foram comparadas quatro estratégias de análise de dados: análise de variância (ANOVA) de mensurações repetidas (pré e pós-teste), ANOVA com escores de ganho, ANOVA com categorização dos sujeitos no pré-teste (*blocking*) e análise de covariância (ANCOVA). Foram ainda consideradas interpretações alternativas para os efeitos encontrados. Por fim, discutiu-se a adequação de tais estratégias para a avaliação de programas sociais.

Palavras-chave: avaliação de programas sociais; qualificação do trabalhador; discussão metodológica.

Social Programs Evaluation: A Comparison of Data Analysis Strategies

ABSTRACT - This article presents a discussion about methodological questions related to social programs evaluation, taking as an example the case of the National Worker Qualification Plan – PLANFOR. The present study used a quasi-experimental design containing three groups: two experimental groups and one control group. The research was realized from 2001 to 2004, with data collection before and after the program, and comprising 360 organizations, one per city, from all Brazilian regions. For data collection, it was used an instrument with quantitative indicators (social and economic). Four data analysis strategies were compared: analysis of variance (ANOVA) of repeated measurement (pre- and post-test scores), ANOVA with gain scores, ANOVA after subjects' categorization on their pre-test (blocking), and analysis of covariance (ANCOVA). Alternative explanations of the observed effects were considered. Finally, the suitability of these strategies to the evaluation of social programs was discussed.

Keywords: social program evaluation; worker qualification; methodological discussion.

O Estado tem determinadas funções sociais, com destaque, sobretudo, para aquelas relativas à educação, saúde, segurança, proteção aos desamparados e responsabilidade na geração de condições que impeçam a manutenção e o crescimento da pobreza. A experiência dos países desenvolvidos ensina que nem a sociedade civil e nem os mercados foram capazes de reduzir a pobreza. O ator principal da promoção de uma sociedade mais justa nos países do primeiro mundo, sobretudo a partir da Segunda Guerra Mundial, foi o Estado. Para cumprir suas funções sociais, o Estado estabelece políticas públicas que são executadas por meio de projetos e programas, para os quais são estabelecidos objetivos e metas. Mas será que esses objetivos e metas estão sendo atingidos? Qual tipo de delineamento e de análise de dados permite melhor avaliar os resultados dos programas sociais?

1 Endereço: Estrada de Itacoatiara, 101. Cond. Jardim Itacoatiara, Casa 04. Niterói - RJ. CEP 24348-000. *E-mail*: luciana.mourao@pq.cnpq.br.

É inegável que tem havido no mundo, de forma geral, e nos países em desenvolvimento, em especial, aumento das ações e políticas sociais, a partir de uma demanda gerada pelo agravamento das desigualdades sociais. No Brasil, os programas sociais têm proliferado e vem-se tentando construir uma cultura de avaliação das políticas e dos programas sociais. As avaliações de programas permitem desenvolver e fortalecer a ação pública, com dois objetivos principais: oferecer subsídios para melhoria da eficiência e eficácia da administração desses programas, e oferecer respostas à sociedade sobre a efetividade social das políticas implantadas, em um processo de prestação de contas. Porém, embora a importância da avaliação dos programas sociais seja praticamente um consenso, ainda há muitos desafios em países como o Brasil, que há pouco mais de duas décadas começou de fato a avaliar os seus programas sociais de maneira sistemática. Ainda há muito a se aperfeiçoar em termos de delineamentos, métodos de pesquisa e sistematização de procedimentos que permitam uma leitura mais fidedigna da efetividade das políticas e dos programas sociais. Assim, o

objetivo do presente artigo é apresentar uma discussão sobre métodos de avaliação de programas, tomando como exemplo um programa educacional avaliado a partir de um estudo quase-experimental.

O programa público cuja avaliação é relatada neste artigo é o Plano Nacional de Qualificação do Trabalhador – PLANFOR, em parceria com a Federação Nacional dos Pais e Amigos dos Excepcionais – Fenapae, que é um programa de formação profissional que visa a geração de emprego e renda para pessoas excluídas do mundo do trabalho. O delineamento da pesquisa foi de um estudo quase-experimental, com três grupos: dois experimentais e um grupo controle. A pesquisa foi realizada durante três anos, com coleta de dados antes e depois da realização da intervenção (os cursos de formação profissional). A seguir será apresentado um referencial teórico de avaliação de programas, uma breve discussão sobre a avaliação de programas públicos no Brasil, o método utilizado na presente pesquisa e os resultados encontrados sobre o aproveitamento do investimento social realizado no caso do PLANFOR.

Avaliação de Programas Públicos

Oskamp (1981) relata uma ampla revisão realizada nos Estados Unidos, há quase três décadas, que contemplou a avaliação dos efeitos de um grande número de programas públicos de inovações médicas e sociais. A revisão permitiu concluir que somente 20% dos programas analisados produziram efeitos claros e positivos. Essa conclusão alertou para a necessidade de pesquisas de avaliação e, especificamente, de avaliação de impacto, uma vez que esse tipo de avaliação é o que permite de forma mais direta mensurar os efeitos dos programas sociais.

As análises de impacto permitem determinar se um programa surtiu os efeitos propostos. Rossi e Freeman (1989) consideram que a avaliação de impacto é de suma importância para os programas sociais e que elas são especialmente recomendadas quando: (a) os debates políticos giram em torno das probabilidades de eficácia de uma política ou programa proposto; (b) é necessário provar a forma mais efetiva de desenvolver e integrar os diferentes elementos do programa; (c) os programas puderem sofrer modificações, sejam elas radicais ou apenas de ajustes.

Na América Latina, inclusive pela necessidade de obtenção de recursos externos, tem sido crescente a preocupação com a avaliação de programas sociais, o que pode ser indicado, pela aparição, na década de 1990, de vários trabalhos e documentos preparados por autores latino-americanos, os quais sistematizam técnicas, elaboram criticamente alguns dos principais enfoques da avaliação e relatam experiências de avaliação de programas públicos nesses países.

Para Sulbrandt (1993), a característica dos trabalhos latino-americanos até o início da década de 90 vinha sendo uma apresentação fundamentalmente metodológica e epistemológica, e uma explicação das várias maneiras possíveis de realizar avaliações, sem relação com os conteúdos substantivos dos programas nem com as formas de implementá-los. Para o autor, embora o trabalho de sistematização de enfoques e técnicas disponíveis fosse de qualidade, essas apresentações não realizavam uma discussão sobre a validade e a utilidade prática dos distintos enfoques de avaliação como acontece nos países mais avançados em termos de avaliação de programas. O autor alerta para as críticas que têm sido feitas à fragilidade metodológica das avaliações – o que é particularmente grave na América Latina. Oskamp (1981), porém, analisa que mesmo nos países mais avançados, embora as avaliações de programas públicos e privados sejam feitas há mais tempo, somente a partir da década de 70 a pesquisa de avaliação tornou-se uma disciplina profissional.

No início da década de 80, Oskamp (1981) enfatizou a necessidade de se adotar uma concepção de ciência nas pesquisas de avaliação, pois a aplicação dos resultados da pesquisa, a teoria da decisão ou o sistema gerencial recebiam maior atenção do que outros conteúdos como o delineamento e a validade da pesquisa. Ao mesmo tempo, Oskamp alertou para o perigo da relevância exclusiva nos resultados se um avaliador optasse por adotar um excesso de rigidez, uma abordagem, como ele denominou, "ultracientífica".

Mais de duas décadas depois, a preocupação com o rigor e com o método na realização das avaliações de programas permanece como uma questão central. Autores estrangeiros como Cronbach (1983), Rossi e Freeman (1988), Posavac e Carey (1997), Shadish, Cook e Campbell (2002), Wholey, Hatry e Newcomer (2004), Worthen, Sanders e Fitzpatrick (1997/2004), e autores brasileiros como Carvalho (2003), Murta, Laros e Tróccoli (2005) e Hedler (2007) apontam a importância da questão metodológica, do rigor do delineamento, do uso correto das medidas e da discussão da validade das pesquisas realizadas nos processos de avaliação de programas. Apesar dos significativos avanços na área, sobretudo nos países mais desenvolvidos, a avaliação de programas ainda apresenta uma série de questões que carecem de debate mais aprofundado pelos cientistas e pelos profissionais que atuam na área.

Sulbrandt (1993) alerta que mesmo quando são realizadas com rigor metodológico, como é feito nos Estados Unidos, na França e na Alemanha, as avaliações são alvo de severas críticas que têm gerado uma crise de identidade para os avaliadores profissionais. A crítica mais forte tem sido a de que os resultados não são utilizados para a tomada de decisões, demonstrando, em última instância, que as avaliações são irrelevantes.

Oskamp, ainda em 1981, alertou para o fato de que a todo o momento era lançado um novo programa social e que a população queria saber se o programa estava funcionando ou não. Para responder a essa questão, era necessário realizar pesquisas de avaliação. Mas, de que adiantaria realizá-las e comunicar os resultados à comunidade se os mesmos não são usados nos processos de tomada de decisão pelos gestores dos programas públicos?

De fato, é preciso discutir quanto e quão bem as pesquisas de avaliação são usadas e não apenas os seus aspectos metodológicos e de validade. Oskamp (1981) considerava que a pesquisa de avaliação era apenas um dos muitos fatores que poderiam influenciar as decisões políticas, e que ela, freqüentemente, tinha efeitos graduais e cumulativos no curso dos programas sociais. O alerta feito por Oskamp continua sendo válido atualmente, embora esteja sendo criado um consenso

de que é preciso dedicar maior atenção ao resultado das avaliações e tenha aumentado significativamente o esforço para avaliar a processo, o impacto e a efetividade de programas públicos ou privados.

Uma questão que merece cuidado em relação à avaliação de programas é o fato de que a maior parte deles produz resultados em médio e longo prazo, o que demanda delineamentos de avaliação que incluam estudos longitudinais. Porém, Cronbach (1983) adverte que um dos maiores problemas da avaliação é que os programas mudam de um ano para outro e que raramente uma linha de ação mantém-se constante por um período que permita aos avaliadores mensurar resultados em curto, médio e longo prazo, sem a intervenção de variáveis que foram modificadas durante o processo de avaliação.

Para Rossi e Freeman (1989), duas possibilidades básicas de delineamento servem como um marco de referência comum a todas as avaliações de impacto: os grupos experimentais devem ser comparados tão rigorosamente quanto possível aos grupos controle e, na ausência de grupos controle, a informação sobre o comportamento pós-intervenção pode ser comparada com as medições de um período anterior ou com os resultados e conjecturas que teriam ocorrido sem a intervenção. Os autores explicam que o delineamento experimental ou quase-experimental depende da comparação de um ou mais grupos experimentais com um grupo controle. Ainda que muitas análises de impacto não possam usar uma técnica estritamente experimental, todos os desenhos de análises de impacto comparam os resultados da intervenção com alguma estimativa do que tem ocorrido ou com o que vá ocorrer na ausência da intervenção.

Planejar o delineamento a ser adotado na avaliação do programa é uma das etapas mais fundamentais, pois um erro de delineamento pode anular todo um esforço despendido na realização da pesquisa. Não há prescrições para a melhor abordagem metodológica ou técnica de coleta de dados. O foco deve estar na seleção do método que responda da melhor forma possível às questões de avaliação dos tomadores de decisão, considerando os recursos e o tempo disponível. Assim, a ênfase deve estar na credibilidade da abordagem e das técnicas escolhidas e no quanto elas são apropriadas para responder às principais questões da avaliação de impacto.

Carmines e Zeller (1979) definem mensuração como um processo de associar conceitos abstratos a indicadores observáveis, envolvendo o plano para classificação explícita e organizada dos dados em termos das questões centrais que estão na mente do pesquisador. Os métodos quantitativos são usados extensivamente em *surveys* que usam instrumentos padronizados, o que aumenta a relação custo-benefício no processo de análise dos dados. A escolha da abordagem, se qualitativa ou se quantitativa, deve levar em conta o fato de que cada uma tem seu uso específico e que a escolha depende das questões a serem avaliadas. De fato, o processo de avaliação de programas é bastante complexo e demanda não só a compreensão do arcabouço teórico que dá sustentação às políticas públicas, mas também a utilização adequada de metodologia e delineamento de pesquisa.

Assim, tanto as metodologias quantitativas como as qualitativas apresentam importantes contribuições para os processos de avaliação de programas públicos. Uma não substitui a outra e não é desejável que os delineamentos sejam feitos considerando-se apenas uma delas. As limitações de recursos e prazo muitas vezes impedem que os delineamentos sejam traçados como deveriam, mas é preciso cuidar para que não se opte por avaliações que pouco contribuem para aquele e para programas futuros.

Nesse sentido, Uchimura e Bosi (2002) avaliam que a dimensão subjetiva da qualidade de programas públicos "guarda, certamente, em seus meandros, muitos aspectos a serem desvelados, já que pertence ao plano das nuanças, do profundo, do particular" (2002, p. 1567). As autoras avaliam que é premente considerar a participação dos atores e, principalmente, dos usuários, na avaliação. De fato, essa é uma questão que precisa ser considerada, porque em última instância, os programas são desenhados em função de seus beneficiários.

Avaliação de Programas Públicos no Brasil

O momento atual é privilegiado para o debate sobre intervenção social no Brasil por cinco motivos: (a) permanência das discrepâncias sociais e da má distribuição de renda no Brasil; (b) crescente aumento da consciência da população em relação às questões sociais; (c) visível atuação do governo federal em programas sociais; (d) tentativa de retomada do desenvolvimento e de geração de emprego; e (e) necessidade de atender a segmentos tradicionalmente marginalizados – como é o caso das pessoas com deficiência mental. Sobre essa demanda por uma cultura de avaliação de programas sociais, é importante considerar que os problemas no Brasil e na América Latina são essencialmente sociais, econômicos e políticos, principalmente de desigualdade de renda, pobreza e exclusão social (Barreira, 2002).

No Brasil, problemas esperando por políticas e programas públicos abundam – apenas para citar alguns, poder-se-ia falar naqueles relativos à segurança, moradia, educação, saúde e a trabalho. Mas as soluções não são simples. Não basta apenas conceber programas sociais. É preciso ter recursos para financiá-los e pensar em soluções que não sejam meramente assistencialistas e que ajudem o País a construir um processo de desenvolvimento sustentável. Muitas questões estão envolvidas no processo de tomada de decisão acerca da escolha das políticas e dos programas públicos, entre elas: (a) Quais são os problemas mais graves e que devem merecer maior atenção?; (b) Quais são os públicos-alvo prioritários?; (c) Que questões regionais demandam intervenção federal?; (d) Com que recursos o governo irá gerir os programas? Quando essas perguntas conseguem ser respondidas e o Estado desenvolve estratégias para implementar os programas, surgem outras perguntas com destaque para pelo menos sete: (a) O programa contempla as necessidades das pessoas que serão servidas?; (b) Qual o grau de participação dos principais interessados?; (c) O programa prevê todos os planos de implementação?; (d) Os objetivos são factíveis e foram estabelecidos de maneira clara?; (e) Há suporte teórico?; (f) O programa é aceito pela sociedade e pelos demais *stakeholders* envolvidos?; (g) Quais são os mecanismos públicos e privados de controle da eficiência, da eficácia e da efetividade dos mesmos?

Para atender a esta última questão é necessário que sejam desenvolvidos processos de avaliação de programas. O Brasil tem pouca tradição em avaliação de programas públicos, contudo, a avaliação de programas tem evoluído bastante, sobretudo a partir da década de 80, quando fica mais evidente a emergência e luta de movimentos sociais populares urbanos (Barreira, 2002). Chianca e Yourker (2004) apresentam quatro razões para essa evolução: (a) a criação de organizações de avaliação profissional; (b) a intensificação da difusão das idéias e do uso da avaliação profissional em três setores chave – o estatal, o privado e o filantrópico; (c) o aumento do número de publicações associadas à avaliação; (d) o estabelecimento de programas de capacitação sobre avaliação, em curto prazo, para graduados.

Entre as formas de controle público, o Tribunal de Contas da União – TCU, desde 1990, vem desenvolvendo trabalhos, sob a modalidade de auditorias operacionais, que tratam não só da questão financeiro-contábil, mas da questão da eficiência, da economicidade, da eficácia e da efetividade da gestão pública. O Tribunal desenvolveu um programa de capacitação de seus técnicos com o objetivo de prover o TCU dos meios necessários à prática de avaliação de programas públicos orientada a resultados (Bugarin, 2002).

Assim, o TCU ampliou sua atuação, deixando de considerar apenas os aspectos de legalidade e conformidade, passando a desempenhar os trabalhos de avaliação de desempenho que envolvem questões de causa e efeito bem mais complexas. Inicialmente, o TCU escolheu 11 projetos públicos para desenvolver o exercício completo de avaliação de programas, como forma de possibilitar um exercício prático ao corpo técnico do Tribunal. Formalizou parcerias com órgãos da administração pública, entidades acadêmicas e de natureza privada que desenvolveram trabalhos sobre avaliação de programas públicos, a fim de propiciar a troca e o acesso a informações, experiências e conhecimentos. Desde então, o TCU desenvolveu uma tecnologia própria de avaliação e criou uma secretaria específica para as auditorias de natureza operacional que são aquelas nas quais se encontram as avaliações dos programas públicos.

Hedler (2007) realizou uma meta-avaliação da auditoria de natureza operacional do TCU, de seus critérios e padrões, verificando a validade e a confiabilidade dos procedimentos e dos resultados. A autora analisou auditorias das áreas de agricultura, assistência social, cidadania, comércio e serviços, educação, energia, habitação, meio ambiente, saúde e trabalho. Hedler utilizou a abordagem qualitativa para analisar 10 relatórios de avaliações de programas públicos realizadas por equipes do TCU e adotou os critérios do *Joint Committee* para a realização da análise. A autora chegou à conclusão de que o processo de auditoria de programas sociais está associado a variáveis como contexto político, características dos programas sociais, e enfoque, métodos e técnicas das auditorias. A autora descreve uma série de dificuldades metodológicas para a realização das auditorias, muitas das quais podem ser estendidas para avaliações de programas sociais realizadas fora do contexto de auditorias públicas. Entre as dificuldades internas do TCU, Hedler aponta: mortalidade instrumental, prazo para a auditoria, dificuldades técnicas da auditoria e formação do grupo de contato de auditoria.

Além do TCU, o Brasil conta com outras instituições que têm tecnologia na área, tais como: Fundação Getulio Vargas, Universidade de Brasília, Universidade de Campinas, Universidade de São Paulo, Ministério da Educação, Ministério do Planejamento e Orçamento, Secretaria Federal de Controle do Ministério da Fazenda e outros da iniciativa privada, que já trabalham, de longa data, com o tema em questão (Bugarin, 2002).

Em outubro de 2001, foi realizado um seminário internacional sobre o tema, denominado de *Avaliação de Programas: Experiências da Austrália, Canadá e Brasil*. Os casos da Austrália e do Canadá são interessantes porque esses países são referências na avaliação de programas públicos. A realização desse seminário ocorreu em função da reestruturação das atividades do governo federal em programas no Plano Plurianual 2000-2003 Avança Brasil. O objetivo dos organizadores era que a troca de experiências propiciasse melhorias nos programas e nos processos de avaliação e tomada de decisão relativos aos mesmos. De acordo com dados apresentados no seminário, na Austrália, resultados da avaliação influenciaram 77% das novas propostas de política e 65% das decisões de cortes no processo orçamentário de 1994-1995. A avaliação foi integrada sistematicamente na gestão e planejamento corporativo e de programas em meados da década de 80. Todos os programas são avaliados a cada três a cinco anos; as propostas de política incluem uma estratégia de avaliação e os resultados das avaliações são divulgados ao público. O Brasil ainda está longe de ter sistemas de avaliação tão controlados e eficientes, mas tem buscado participar do debate sobre o tema e repensar as questões centrais sobre avaliação de programas.

Antunes (2001) explica que a avaliação pode ser conceituada como um processo que tem como finalidade a determinação sistemática e objetiva da relevância, efetividade e impacto de programas ou políticas públicas, à luz de seus próprios objetivos. Porém, nem sempre os programas apresentam objetivos claros e mensuráveis. Além disso, o processo de avaliação de programas no Brasil precisa levar em conta a questão dos delineamentos. Ao contrário de países que adotam experimentos e quase-experimentos para a avaliação de praticamente todos os programas, aqui, a adoção de delineamentos com grupos de controle e ênfase no rigor metodológico são raros. Cano (2002) considera que a qualidade da avaliação de um programa depende tanto ou mais do delineamento da pesquisa do que da natureza das medidas empregadas. O autor avalia que a insistência nas questões de delineamento de pesquisa para garantir a comparabilidade dos grupos, o controle de variáveis intervenientes e a superação dos problemas de seleção diferencial são cruciais para poder avaliar uma intervenção social.

Em outro estudo, Carvalho (2003) recuperou as avaliações de programas sociais desenvolvidas pela Fundação Seade, com ênfase nos aspectos metodológicos. Para a autora, as avaliações das ações públicas de provisão habitacional, intervenção em cortiços, oferta de educação supletiva à distância, estágio para estudantes de ensino médio e atividades culturais para populações socialmente vulneráveis, mostram singularidades e regularidades que se somam às reflexões sobre metodologias de avaliação, em particular no campo

das políticas sociais e sobre o lugar da avaliação no processo das políticas públicas. Carvalho concluiu que há fragilidade nos critérios utilizados no registro e armazenamento dos dados dos programas públicos, dificultando e, por vezes, impossibilitando seu acompanhamento e avaliação. A autora também concluiu que a atividade de avaliação se realiza pela utilização de diferentes métodos da pesquisa social, que foi o que ocorreu na avaliação dos cinco programas que ela revistou. Por fim, Carvalho apontou uma última questão sobre a posição ocupada pela avaliação na fronteira entre a produção de conhecimentos e a sua aplicação prática para o aperfeiçoamento do sistema das políticas públicas. Para ela, alguns estudos sugerem o "aprofundamento das noções de eficiência, eficácia e efetividade na avaliação de políticas públicas e, de forma correlata, de resultados, efeitos, mudanças e impactos" (p.195).

Por sua vez, Campêlo (2006) afirma que a avaliação de impacto apresenta-se como um instrumento de alta complexidade e que sua operacionalização passa pela capacidade de identificar as dimensões existentes no contexto a ser avaliado. Ou seja, quais são os indicadores que de fato apontam que houve impacto do programa.

Método

O presente estudo contemplou a avaliação de um programa de formação profissional que visa capacitar professores para prepararem pessoas com deficiência mental para o mundo do trabalho. O programa, resultante de uma parceria da Federação Nacional das APAEs – Fenapae e do Ministério do Trabalho e Emprego, contemplou treinamentos do Projeto de Formação de Formadores em Educação Profissional - PLANFOR, realizados no ano de 2001. De maneira sintética, o problema de pesquisa pode ser assim descrito: Qual o grau de êxito do programa de formação profissional do PLANFOR, considerando como indicadores a mudança organizacional e a geração de empregos para os deficientes mentais?

A pesquisa foi realizada durante os anos de 2001 a 2004, com coleta de dados antes e depois da realização da intervenção (o programa de treinamento). A avaliação do impacto na mudança organizacional foi feita com a utilização de indicadores objetivos: aumento no número de oficinas de formação profissional e no número de aprendizes. Também foram avaliados os resultados relativos ao aumento no número de pessoas com deficiência mental empregadas. O delineamento também incluiu a análise de possíveis explicações alternativas para os efeitos observados (Produto Interno Bruto - PIB, taxa básica de juros, taxa de desemprego e mudanças na legislação) que poderiam causar mudanças nos resultados independentemente da realização do treinamento. A inclusão da análise dessas variáveis é importante porque o aumento do número de pessoas empregadas responde a muitos outros fatores além do programa de treinamento.

Assim, o impacto foi mensurado por meio de instrumentos quantitativos. Esses dados foram informados por presidentes ou diretores das APAEs por meio de correspondência enviada pelo correio. Foi assinado o termo de consentimento livre e esclarecido pelos representantes dessas instituições.

O estudo teve delineamento quase-experimental com 360 APAEs, divididas em três grupos de tratamento do mesmo tamanho (120 APAEs):

a) grupo experimental principal (APAEs que tiveram o coordenador e pelo menos três professores participando do treinamento);
b) grupo experimental secundário (APAEs que tiveram apenas o coordenador treinado);
c) grupo controle (APAEs que não tiveram participantes no treinamento).

Os grupos foram definidos dessa forma porque havia interesse em saber não só se a participação no programa geraria algum tipo de impacto, mas também se o número de profissionais capacitados contribuiria para explicar os resultados da ação de treinamento. Considerando-se esses três grupos, trabalhou-se com o universo das APAEs que participaram e que enviaram as informações e não com uma amostra das mesmas. Porém, como o critério definido foi a participação no programa, em função de ser este o interesse principal do estudo, algumas outras variáveis relativas a essas APAEs poderiam apresentar distribuição heterogênea como será discutido nos resultados.

Segundo Posavac e Carey (2003), as qualidades de um bom processo de avaliação estão fundamentadas no uso de medidas que apresentem: multiplicidade de fontes e de variáveis, não-reatividade (a medição não pode alterar as situações em análise), validade (mensuração daquilo que o programa está desenhado para mudar), confiabilidade (diferentes observadores do mesmo fenômeno reportam níveis similares das variáveis em análise), sensibilidade para captar mudanças pequenas e efetividade de custo-benefício. Assim, o delineamento da presente pesquisa buscou considerar esses aspectos e apresenta as seguintes características: estudo longitudinal, delineamento quase-experimental, estudo de explicações alternativas para os efeitos encontrados, utilização de dados primários e de dados secundários, opção pela avaliação somativa (análise de resultados) e múltiplas técnicas de análise de dados.

Resultados e Discussão

Para a análise da ocorrência de impacto do programa avaliado foram utilizados quatro procedimentos: ANOVA de mensurações repetidas, ANOVA com escores de ganho, ANCOVA e ANOVA depois da categorização dos sujeitos no pré-teste (*blocking*). A seguir serão apresentados um a um os procedimentos e a discussão dos resultados dos mesmos e da conveniência de utilizá-los.

ANOVA de mensurações repetidas

A ANOVA com utilização dos níveis fatoriais de mensurações repetidas só pode ser utilizada quando há dados do pré e do pós-teste pareados. Antes de se realizar esse tipo de procedimento de análise é preciso se certificar do grau de fidedignidade dos valores do pré e pós-teste. Como os valores medidos referem-se a indicadores objetivos e não perceptuais, não é possível calcular o grau de

fidedignidade do pré e nem do pós-teste, sendo possível apenas estimar tal fidedignidade.

Os dados coletados referem-se ao número de oficinas e de aprendizes em educação profissional em cada APAE. Acredita-se que esse dado tenha alta fidedignidade porque os aprendizes são registrados e esses números são controlados pelo Ministério da Educação, não podendo haver divergência entre o número de alunos e o de registros. O mesmo vale para as oficinas, pois cada oficina aberta em uma APAE é registrada na mesma, e informada para a Delegacia Regional na qual a APAE está inserida. Porém, a fidedignidade de 100% é bastante improvável, pois, mesmo nos casos de indicadores controlados por fontes externas (como é o caso do número de aprendizes e de oficinas de educação profissional nas APAEs), é possível que ocorram erros de digitação no momento de informar tais dados. Assim, estima-se um grau de confiabilidade da ordem de 95%.

Para realizar a ANOVA a partir dos escores do pré e do pós-teste, o banco de dados foi dividido nos três grupos de tratamento. Foi realizada uma ANOVA para cada variável dependente (número de oficinas de educação profissional nas APAEs, número de aprendizes e número de pessoas com deficiência mental empregadas). A Tabela 1 mostra as médias e os erros-padrão das três variáveis dependentes (VDs) da pesquisa realizada.

Os resultados da Tabela 1 indicam, em primeiro lugar, que os escores de pós-teste para as três VDs são mais altos que os escores do pré-teste para os três grupos da pesquisa: o grupo controle e os dois grupos experimentais. Em segundo lugar, os dados sugerem que as diferenças entre os escores de pré-teste e pós-teste tendem a ser maiores para o grupo experimental principal em comparação com os demais grupos. Em terceiro lugar, os dados sugerem que o grupo controle e os dois grupos experimentais não foram equivalentes antes da intervenção uma vez que os escores no pré-teste são diferentes para os três grupos de pesquisa. Essa observação é de suma importância uma vez que a não-equivalência de grupos impossibilita a simples comparação dos escores no pós-teste dos três grupos da pesquisa. Para verificar se as diferenças dos escores do pré-teste são realmente estatisticamente significativas, foram calculados os intervalos de confiança de 95% (IC de 95%) dos escores do pré-teste para as três VDs. A Tabela 2 mostra esses intervalos de confiança.

Os resultados da Tabela 2 indicam que para a primeira VD, número de oficinas, há diferenças significativas entre o grupo experimental principal e os demais grupos da pesquisa. As diferenças dos escores no pré-teste do grupo controle e do grupo experimental secundário não são significativas. Para a segunda VD, número de aprendizes, os ICs de 95% mostram que todos os grupos têm escores estatisticamente diferentes do pré-teste. Finalmente, os ICs de 95% da terceira VD, número de pessoas empregadas, mostram que não há diferenças significativas entre os escores do pré-teste dos três grupos. Resumindo, pode-se concluir que em relação à primeira e à segunda VD, os três grupos não são equivalentes. A falta de equivalência entre grupos impossibilita a execução de uma ANOVA simples sem antes fazer uma correção para as diferenças dos grupos da pesquisa dos escores no pré-teste (Posavac & Carey, 2003). Uma alternativa é a utilização de uma ANOVA com escores de ganho, que é o que será apresentado a seguir.

ANOVA com escores de ganho

Uma ANOVA com escores de ganho considera os escores resultantes da diferença entre o pré-teste e o pós-teste. Uma questão a ser considerada é que a fidedignidade de um escore de ganho é, em geral, consideravelmente menor do que a fidedignidade do pré ou pós-teste. Quanto a fidedignidade vai diminuir depende da correlação entre o pré e o pós-teste: quanto mais alto esta correlação, menor a fidedignidade do escore de ganho.

Assim, para cada variável dependente, foi calculada a fidedignidade do escore de ganho. Os índices de fidedignidade para os escores de ganho das três VDs (número de oficinas, número de aprendizes e número de aprendizes que conseguiram um emprego) são, respectivamente, 0,72, 0,74 e 0,79. Os valores da fidedignidade dos escores de ganho – relativamente altos – aumentam a confiança nos resultados baseados em escores de ganho. A Tabela 3 mostra as médias e erros-padrão dos escores de ganho das três VDs deste estudo.

Os resultados da Tabela 3 sinalizam que todos os três grupos da pesquisa mostram escores de ganho positivos e estatisticamente significativos. Merece destaque nessa tabela que há crescimento dos escores de ganho do grupo controle para o grupo experimental secundário e deste para o grupo experimental principal. O grupo experimental principal mostra escores de ganho mais altos para todas as três VDs. O que os dados não mostram é que existem diferenças significativas nos escores de ganho entre os grupos. Para resolver essa questão foi executada uma ANOVA das diferenças entre os grupos dos escores de ganho, conforme indica a Tabela 4.

Os resultados da Tabela 4 mostram que, para as três VDs, houve diferenças significativas entre os grupos (os dois experimentais e o controle) em relação aos escores de ganho. Vale salientar que os resultados revelam que existem diferenças significativas entre os três grupos, mais não indicam entre quais grupos. Por exemplo, a diferença do grupo controle e grupo experimental secundário nos escores de ganho na primeira VD (número de oficinas) pode não ser significativa.

Em relação ao uso da ANOVA com escores de ganho é importante salientar que, segundo Brace, Kemp e Snelgar (2003), o objetivo de remover o efeito dos escores do pré-teste nos escores de ganho pode não ser alcançado. Para saber se o efeito do pré-teste foi ou não anulado, é preciso verificar as correlações entre os escores de ganho e os escores do pré-teste, que no caso da presente pesquisa estavam baixos (0,04, 0,10 e 0,15, respectivamente, para as três VDs). A correlação baixa encontrada neste estudo indica que os escores de ganho conseguiram reduzir consideravelmente a influência inicial dos escores antes da intervenção.

Dugard e Todman (1995) avaliam, todavia, que, mesmo nos casos em que não há problemas de correlação alta entre os escores de ganho e os escores iniciais do pré-teste, a ANCOVA é o procedimento de análise de dados mais adequado. Para os autores, a preferência pela ANOVA decorre mais do costume de alguns pesquisadores com a mesma que de argumentos baseados em resultados melhores com a opção pela ANOVA. Em seguida serão mostrados os resultados com ANCOVA em que os escores de pré-teste são tratados como covariantes.

Tabela 1. Médias (M) e erros-padrão da média (EP) das três variáveis dependentes: número de oficinas de formação profissional, número de aprendizes e número de pessoas com deficiência mental empregadas (pré e pós-teste).

Grupos da Pesquisa	N	Número de oficinas				Número de aprendizes				Número de empregados			
		Pré-teste		Pós-teste		Pré-teste		Pós-teste		Pré-teste		Pós-teste	
		M	EP	M	EP	M	EP	M	EP	M	EP	M	EP
Grupo Controle	118	3,2	0,29	3,6	0,35	30,6	2,52	31,7	2,73	4,1	0,71	4,7	0,75
Grupo Exp. Secundário	119	2,6	0,23	3,1	0,27	21,7	1,75	24,8	1,84	2,9	0,51	4,8	0,68
Grupo Exp. Principal	120	4,8	0,31	5,9	0,40	49,6	3,49	58,6	4,23	4,2	0,54	7,4	0,86

Tabela 2. Intervalos de confiança de 95% dos escores no pré-teste nas três variáveis dependentes (VDs) para os diferentes grupos de pesquisa.

Grupos da Pesquisa	Média	Erro Padrão	Intervalo de Confiança de 95%	
			Limite Inferior	Limite Superior
	VD I: Número de oficinas de formação profissional no pré-teste			
Grupo Controle	3,2	0,29	2,6	3,8
Grupo Experimental Secundário	2,6	0,23	2,1	3,1
Grupo Experimental Principal	4,8	0,31	4,2	5,4
	VD II: Número de pessoas com deficiência mental aprendizes no pré-teste			
Grupo Controle	30,6	2,52	25,6	35,6
Grupo Experimental Secundário	21,7	1,75	18,2	25,2
Grupo Experimental Principal	49,6	3,49	42,6	56,6
	VD III: Número de pessoas com defic. mental empregadas no pré-teste			
Grupo Controle	4,1	0,71	2,7	5,5
Grupo Experimental Secundário	2,9	0,51	1,9	3,9
Grupo Experimental Principal	4,2	0,54	3,1	5,3

Tabela 3. Médias e erros-padrão (EP) dos escores de ganho das três variáveis dependentes.

Grupos da Pesquisa	N	Número de oficinas		Número de aprendizes		Número de empregados	
		Média	EP	Média	EP	Média	EP
Grupo de Controle	118	0,4	0,17	1,1	1,39	0,6	0,36
Grupo Exp. Secundário	119	0,6	0,18	3,1	1,20	1,9	0,41
Grupo Exp. Principal	120	1,1	0,26	9,0	2,31	3,2	0,61

Tabela 4. ANOVA das diferenças entre pré e pós-teste (escores de ganho) para as três variáveis dependentes.

Escores de ganho		Soma dos quadrados	Graus de liberdade	Média dos quadrados	F	Significância
Diferença do número de oficinas	Entre grupos	40,5	2	20,2	4,1	0,02
	Nos grupos	1.788,0	354	5,0		
	Total	1.808,5	356			
Diferença do número de pessoas com deficiência mental aprendizes	Entre grupos	4.092,5	2	2.046,2	5,9	< 0,01
	Nos grupos	123.463,8	354	348,8		
	Total	127.556,3	356			
Diferença do número de pessoas com deficiência mental empregadas	Entre grupos	406,8	2	203,4	7,6	< 0,01
	Nos grupos	9.459,3	354	26,7		
	Total	9.866,1	356			

Tabela 5. ANCOVA: Teste de homogeneidade das inclinações da regressão para as três variáveis dependentes (VDs).

Fonte	Soma dos quadrados	Graus de liberdade	Média dos quadrados	F	Significância
	VD I: Número de oficinas de formação profissional no pós-teste				
Modelo corrigido	142,4	5	28,5	155,5	< 0,01
Intercepto	5,1	1	5,1	27,6	< 0,01
Grupo	1,7	2	0,8	4,6	0,01
Grupo * Nº Oficinas no pré-teste	1,3	2	0,7	3,6	0,03
Número de Oficinas no pré-teste	110,1	1	110,1	601,3	< 0,01
Erro	64,3	351	0,2		
Total	1.851,0	357			
Total corrigido	206,6	356	$R^2 = 0,69$ (R^2 ajustado = 0,68)		
	VD II: Número de pessoas com defic. mental aprendizes no pós-teste				
Modelo corrigido	1.708,9	5	341,8	146,4	< 0,01
Intercepto	81,0	1	81,0	35,1	< 0,01
Grupo	20,9	2	10,5	4,5	< 0,01
Grupo *Nº Aprendizes no pré-teste	19,6	2	9,8	4,2	0,02
Número de Aprendizes no pré-teste	1.131,9	1	1.131,9	484,9	< 0,01
Erro	819,3	351	2,3		
Total	14.071,0	357			
Total corrigido	2.528,2	356	$R^2 = 0,68$ (R^2 ajustado = 0,67)		
	VD III: Número de pessoas com defic. mental empregadas no pós-teste				
Modelo corrigido	386,8	3	128,9	173,1	< 0,01
Intercepto	32,3	1	32,3	43,4	< 0,01
Número de empregados no pré-teste	365,6	1	365,6	490,7	< 0,01
Grupo	10,3	2	5,2	6,9	< 0,01
Erro	263,0	353	0,8		
Total	2.369,0	357			
Total corrigido	649,7	356	$R^2 = 0,60$ (R^2 ajustado = 0,59)		

ANCOVA

Segundo Brace e cols. (2003) e Dugard e Todman (1995), para a realização da ANCOVA há os mesmos pressupostos da ANOVA e mais dois outros: a independência entre o tratamento e o covariante (homogeneidade das linhas de regressão) e a mensuração do covariante sem erro. Nesta pesquisa, a mensuração do covariante sem erro foi um pressuposto atendido, pois foram considerados como covariantes o número de oficinas de educação profissional nas APAEs em 2001, o número de aprendizes nas APAEs neste mesmo ano e o número de pessoas com deficiência mental que estavam empregadas antes do início do programa (2001). Conforme explicado anteriormente, esses covariantes são indicadores objetivos, sendo possível apenas estimar tal fidedignidade, porém, pelas razões já expostas, estima-se uma mensuração com grau de fidedignidade da ordem de 0,95.

Quanto à independência entre o tratamento e o covariante, houve violação desse pressuposto para duas das três VDs. Quando os sujeitos são designados randomicamente para o grupo experimental e o grupo controle, e o pré-teste é administrado antes de o tratamento começar, não há o risco de violação do pressuposto. Porém, no caso da presente pesquisa, embora o pré-teste tenha sido administrado antes de a intervenção começar, as APAEs não foram designadas randomicamente para a participação no programa. Para testar a homogeneidade das linhas de regressão (a relação entre a VD e o covariante deve ser similar para os grupos experimentais), o procedimento GLM (*General Linear Model*), do SPSS 11.0, foi executado. Utilizou-se o modelo customizado do GLM *Univariate*, sendo os covariantes os escores do pré-teste, o fator fixo os grupos de pesquisa (tratamento) e incluindo-se no modelo a interação entre eles. Os resultados para as variáveis número de oficinas e número de aprendizes indicaram que a interação entre o tratamento e o pré-teste eram significativas, devendo prevalecer, portanto, a análise do modelo customizado (*custom*) com interação. Já para a variável número de pessoas com deficiência mental empregadas, a interação não foi significativa, optando-se, então, pelo modelo "cheio" (*full factorial*). Os resultados obtidos nesses modelos são mostrados na Tabela 5.

A análise dos resultados da Tabela 5 permite concluir que, para as três variáveis, foram significativos no modelo tanto o tratamento como os escores de pré-teste. Para investigar quais grupos da pesquisa mostram diferenças nas VDs depois de controlar para o efeito do pré-teste (o covariante) foi feito uma análise de comparações *pairwise* dos grupos da pesquisa. A Tabela 6 mostra estas comparações *pairwise*.

A partir das informações da Tabela 6 fica evidente que, para a variável número de oficinas, não se pode considerar significativa a diferença entre o grupo controle e o grupo experimental secundário. Entre o grupo experimental secundário e o grupo experimental primário, há diferença estatisticamente significativa. Para a segunda VD, número de aprendizes, a diferença entre o grupo controle e o grupo experimental principal também não é estatisticamente significativa.

Esses foram os principais resultados da ANCOVA. Porém, como foi encontrada interação em duas das três variáveis dependentes (violando-se, para esses casos, o pressuposto da independência entre o tratamento e o covariante) e autores como Dancey e Reidy (1999/2006) recomendam a utilização do *blocking* quando ocorre esse tipo de violação, foi realizado, também, o procedimento de ANOVA com categorização dos sujeitos no pré-teste (*blocking*).

ANOVA depois da categorização dos sujeitos no pré-teste (blocking)

Os pressupostos para que possa ser realizada a ANOVA com categorização (*blocking*) – homogeneidade da variância nos escores residuais do pós-teste, linearidade na relação pré e pós-teste, e homogeneidade das linhas de regressão – foram atendidos. Para a realização dessa análise, a amostra do pré-teste foi dividida em três grupos para a quantidade de oficinas em 2001 e em outros três grupos para a quantidade de aprendizes neste mesmo ano. A idéia desse procedimento é categorizar a amostra do pré-teste de acordo com a quantidade de oficinas e de aprendizes que cada APAE tinha antes da intervenção, formando grupos de tamanho aproximado. Realizando-se tal procedimento, a amostra, em relação ao número de oficinas, foi dividida em: 0 a 1 oficina (n=109), de 2 a 3 oficinas (n=126) e mais de 3 oficinas (n=122). Porém, ao se realizar a análise do GLM, considerando-se como fatores fixos o grupo e o *block* (categorização do pré-teste), observa-se que, para o número de oficinas, em nenhum dos grupos, o tratamento teria sido significativo, como mostra a Tabela 7.

A leitura dos resultados da Tabela 7 permite concluir que, se os grupos do pré-teste fossem categorizados de acordo com os seus escores iniciais, para nenhum dos três grupos a intervenção teria surtido efeito.

Repetindo-se o mesmo procedimento para a análise do número de aprendizes em cada APAE (Tabela 8), a amostra ficou com os seguintes grupos: até 15 aprendizes (n=120), de 16 a 37 aprendizes (n=118) e mais de 37 aprendizes (n=119). Nesse caso, o tratamento poderia ser considerado significativo para os grupos 1 (até 15 aprendizes na APAE) e 3 (mais de 37 aprendizes na APAE).

Os resultados da Tabela 8 demonstram que não há explicação para o fato de as APAEs do grupo de 16 a 37 aprendizes não obterem resultado com a intervenção como as outras APAEs obtiveram. Poder-se-ia até esperar que as APAEs com menor número de aprendizes apresentassem tal problema, mas não as APAEs com número intermediário. Assim, tanto para a variável número de oficinas como para a variável número de aprendizes, a ANOVA com categorização dos sujeitos no pré-teste (*blocking*) não se mostrou como um método de análise consistente.

Duas razões podem explicar a ineficácia do método: a categorização dos grupos pré-teste foi feita a *posteriori* (após a realização da intervenção) e a pouca variabilidade dos grupos, pois sem a categorização havia 15 tipos de resposta para número de oficinas e 93 para número de aprendizes, e após a categorização, tanto o número de oficinas como o de aprendizes variava apenas entre 1 e 3.

Portanto, do ponto de vista metodológico, seria melhor realizar a ANCOVA, mesmo com a violação do pressuposto de independência entre o tratamento e o covariante, pois a ANOVA com categorização dos sujeitos

Tabela 6. ANCOVA: Comparações *pairwise* para as três variáveis dependentes (VDs).

(I) Grupos da pesquisa	(J) Grupos da pesquisa	Diferença das médias (J – I)	Erro padrão	Significância
VD I: Número de oficinas no pós-teste				
Grupo controle	Grupo experimental secundário	0,03	0,06	0,59
	Grupo experimental principal	0,19	0,06	< 0,01
Grupo experimental principal	Grupo controle	-0,19	0,06	< 0,01
	Grupo experimental secundário	-0,16	0,06	0,01
VD II: Número de pessoas com deficiência mental aprendizes no pós-teste				
Grupo controle	Grupo experimental secundário	0,07	0,21	0,75
	Grupo experimental principal	0,78	0,21	< 0,01
Grupo experimental principal	Grupo controle	-0,78	0,21	< 0,01
	Grupo experimental secundário	-0,72	0,22	< 0,01
VD III: Número de pessoas com deficiência mental empregadas no pós-teste				
Grupo controle	Grupo experimental secundário	0,24*	0,11	0,03
	Grupo experimental principal	0,42*	0,11	< 0,01
Grupo experimental principal	Grupo controle	-0,42*	0,11	< 0,01
	Grupo experimental secundário	-0,17	0,11	0,13

no pré-teste (*blocking*) não se mostrou eficaz para avaliar resultados de delineamentos quase-experimentais com grupo controle. A esse respeito, Dugard e Todman (1995) explicam que a ANCOVA é mais robusta que ANOVA com estratificação quando o pressuposto da linearidade é violado (não foi o caso desta pesquisa, mas é um aspecto que favorece o uso da ANCOVA).

Em relação a presente pesquisa, vale ressaltar que, quando se fala de programas educacionais, como o das APAEs, é comum a violação do atributo de linearidade, pois, em geral, as instituições que partem de situações melhores têm mais condições de aproveitar o programa do que aquelas que se encontram, inicialmente, em situações menos favorecidas. Por exemplo, uma APAE que já tinha certo número de oficinas e de alunos em educação profissional tende a aproveitar mais o treinamento dado aos seus professores que uma APAE que ainda não tinha tal experiência. Como solucionar esse problema é uma questão que merece debate, pois não basta excluir aquelas que se encontram em melhor situação, já que isso significaria penalizar as pessoas com deficiência mental a elas associadas pelo simples fato de estarem em uma instituição com mais recursos.

Finalmente, cabe considerar o que Dugard e Todman (1995) ponderam: a ANCOVA dá ao pré-teste seu *status* correto de covariante em vez de um *status* de uma variável de maior importância em si mesma. A presente pesquisa corrobora os resultados encontrados por esses autores, pois confirma que a ANCOVA tem mais poder, isto é, consegue captar mudanças pequenas não registradas com o uso da ANOVA. Vale ressaltar acerca desses tipos de análises de dados que o presente artigo não pretende esgotar a discussão acerca desses métodos, recomendando-se para pesquisadores em situações de avaliação semelhante a leitura dos autores Dancey e Reidy (1999/2006, especialmente o capítulo 13), Brace, Kemp e Snelgar (2003, especialmente o capítulo 9) e Stevens (2001).

Estudo das explicações alternativas dos efeitos

Embora a ANCOVA tenha sinalizado que os grupos experimentais tiveram melhores resultados que o grupo controle, como a principal variável que está sendo avaliada é a geração de empregos, poder-se-ia atribuir o resultado favorável verificado na pesquisa não só ao programa, mas a outras explicações alternativas. Entre essas explicações poderia estar a redução da taxa básica de juros, o crescimento do PIB, o aumento da taxa de desemprego ou mudanças na legislação que favoreçam a empregabilidade de pessoas com deficiência mental.

Quando há alta taxa de crescimento do PIB, há tendência de aumento de oportunidades de emprego, pois o PIB sinaliza aumento da produção interna do país. Porém, o que se verificou durante os anos da pesquisa foi uma queda do crescimento do PIB. Ou seja, o comportamento do PIB não explicaria o aumento na geração de empregos. A geração de empregos também poderia ser atribuída a um cenário de redução da taxa de juros básica cobrada no país, uma vez que os juros elevados tendem a desaquecer o setor produtivo. Porém, os dados de 2001 a 2003 revelam que a taxa de juros manteve-se em ascensão no período da intervenção, o que tenderia a não reduzir o número de empregos.

Tabela 7. Blocking – Variável dependente: Número de oficinas nas APAEs no pós-teste.

	Fonte	Tipo III soma dos quadrados	Graus de liberdade	Média dos quadrados	F	Significância
Filtro de bloco: Grupo 1 De 0 a 1 oficina no pré-teste	Modelo corrigido	5,8[a]	2	2,9	2,7	0,07
	Intercepto	168,9	1	168,9	155,8	< 0,01
	Grupo	5,8	2	2,9	2,7	0,07
	Block-Oficinas	0,0	0			
	Grupo * *Block*-Of.	0,0	0			
	Erro	114,9	106	1,1		
	Total	319,0	109			
	Total corrigido	120,8	108			
Filtro de bloco: Grupo 2 De 2 a 3 oficinas no pré-teste	Modelo corrigido	8,1[a]	2	4,1	2,3	0,10
	Intercepto	121,2	1	1.211,2	691,3	< 0,01
	Grupo	8,1	2	4,1	2,3	0,10
	Block-Oficinas	0,0	0			
	Grupo * *Block*-Of.	0,0	0			
	Erro	215,5	123	1,8		
	Total	1.437,0	126			
	Total corrigido	223,7	125			
Filtro de bloco: Grupo 3 Mais de 3 oficinas no pré-teste	Modelo corrigido	99,2[a]	2	49,6	2,5	0,09
	Intercepto	5.967,5	1	5.967,5	301,0	< 0,01
	Grupo	99,2	2	49,6	2,5	0,09
	Block-Oficinas	0,0	0			
	Grupo * *Block*-Of.	0,0	0			
	Erro	2.359,5	119	19,8		
	Total	9.950,0	122			
	Total corrigido	2.458,7	121			

Além de analisar o PIB e a taxa de juros, para se aumentar o grau de certeza de que nenhum outro fator estava causando aumento de empregos, analisou-se a própria taxa de desemprego. A série histórica do mês de outubro dos anos de 1991 a 2003 mostra que o desemprego aumentou nesses anos. Portanto, sem a intervenção realizada (a capacitação dos professores), a tendência seria de redução de empregos.

Finalmente, o acompanhamento da legislação sobre emprego para pessoas com deficiência mostrou que *nenhuma* legislação nova foi aprovada entre 2000 e 2004. Essas análises permitem concluir que, de fato, o aumento do número de empregos para pessoas com deficiência deveu-se ao programa de treinamento desenvolvido com recursos do Ministério do Trabalho e Emprego.

Considerações Finais

A importância e a utilização dos processos de avaliação de programas estão em crescimento. Os patrocinadores, os gestores e os executores do programa têm várias razões para realizar avaliação de impacto. Talvez a mais importante dessas razões seja a prestação de contas, uma vez que se trata do uso de recursos públicos e do cumprimento de uma função do Estado. Além disso, a avaliação de programas é uma forma de completar o ciclo, pois em muitos casos a avaliação de impacto é o último passo do programa quando este é concebido como um processo completo.

O propósito da investigação avaliativa é medir os efeitos de um programa em relação às metas e objetivos que ele se propôs alcançar. Assim, a pergunta tradicional que as ava-

Tabela 8. *Blocking* – Variável dependente: Número de aprendizes no pós-teste.

	Fonte	Tipo III Soma dos quadrados	Graus de liberdade	Média dos quadrados	F	Significância
Filtro de bloco: Grupo 1 Até 15 aprendizes no pré-teste	Modelo corrigido	678,6	2	339,3	3,9	0,02
	Intercepto	21.176,4	1	21.176,4	243,2	< 0,01
	Grupo	678,6	2	339,3	3,4	0,02
	Block-Aprendizes	0,0	0			
	Grupo * *Block*-Ap.	0,0	0			
	Erro	10.186,2	117	87,1		
	Total	33.060,0	120			
	Total corrigido	10.864,8	119			
Filtro de bloco: Grupo 2 De 16 a 37 aprendizes no pré-teste	Modelo corrigido	648,8	2	324,4	1,6	0,21
	Intercepto	95.872,4	1	95.872,8	475,0	< 0,01
	Grupo	648,8	2	324,4	1,6	0,21
	Block-Aprendizes	0,0	0			
	Grupo * *Block*-Ap.	0,0	0			
	Erro	23.210,5	115	201,8		
	Total	119.933,0	118			
	Total corrigido	23.859,4	117			
Filtro de bloco: Grupo 3 Mais de 37 aprendizes no pré-teste	Modelo corrigido	22.106,5[a]	2	11.053,2	6,5	< 0,01
	Intercepto	417.994,3	1	417.994,3	245,0	< 0,01
	Grupo	22,106,5	2	11.053,2	6,5	< 0,01
	Block-Aprendizes	0,0	0			
	Grupo * *Block*-Ap.	0,0	0			
	Erro	197.945,0	116	1.706,4		
	Total	858.297,0	119			
	Total corrigido	220.051,4	118			

liações buscam responder é: "Que grau de êxito o programa tem em relação ao atingimento de suas metas?". Mas a avaliação de programas é um tipo de pesquisa aplicada, por isso, é importante não só prestar atenção aos seus aspectos metodológicos, mas também observar se os resultados da avaliação estão afetando o processo de tomada de decisão sobre o programa e as ações, e o conhecimento das pessoas a respeito do mesmo. A não-utilização dos resultados das avaliações tem sido uma das críticas mais freqüentes encontradas na literatura. Nesse sentido, vale destacar que, no presente caso, a avaliação conseguiu trazer algumas contribuições para os gestores: a carga horária dos cursos foi ampliada e decidiu-se não mais enviar apenas os coordenadores para cursos de capacitação, formando sempre uma equipe de profissionais que englobe o coordenador e pelo menos mais três professores.

Outra questão fundamental na avaliação de programas é o delineamento da pesquisa avaliativa. Não há na literatura da área um delineamento que seja apontado como o melhor para todos os casos, mas há severas críticas à utilização de delineamentos pouco científicos. O presente estudo recomenda o uso de experimentos – e quando estes não forem possíveis, de quase-experimentos – com o uso complementar de métodos qualitativos, como o que foi realizado no caso da avaliação do PLANFOR.

Um ponto que está presente no debate sobre avaliação de programas é o da seleção de critérios e de padrões de avaliação. Esse é um aspecto que merece bastante cuidado porque dele dependem todas as outras definições do processo avaliativo. Por um lado, a escolha dos critérios de avaliação varia de acordo com os aspectos do programa, com os objetivos da avaliação e com o tipo de programa que está sendo avaliado. Por outro lado, a revisão de literatura e a falta de tradição em avaliação no Brasil sinalizam para a necessidade de se definir padrões de avaliação e se estabelecer métodos mais rigorosos tanto de delineamento como de análise de pesquisas avaliativas. No caso da pesquisa relatada neste artigo, o critério central estabelecido para a composição dos grupos experimentais e controle foi o grau de participação no programa de treinamento. Porém, outros critérios também foram considerados como resultados anteriores já alcançados por cada uma das APAEs participantes e que poderiam explicar, em parte, os resultados obtidos neste estudo, conforme discussão apresentada.

A presente pesquisa permite concluir que, no caso do Projeto de Formação de Formadores em Educação Profissional da FENAPAE, houve impacto do programa. O quase-experimento sinaliza para resultados mais favoráveis em termos de empregabilidade para as APAEs que enviaram maior número de pessoas para serem capacitadas (grupo experimental principal), seguidas das APAEs que enviaram apenas o coordenador para participar do programa de capacitação (grupo experimental secundário), ficando o grupo que não participou do programa (grupo controle) com os piores resultados de número de alunos empregados. A clareza dos resultados obtidos permite recomendar a utilização de mais estudos quase-experimentais e experimentais que permitam estabelecer relações de causalidade que expliquem o sucesso ou o fracasso dos programas de treinamento e desenvolvimento.

Em relação à técnica de análise de dados, o estudo evidenciou que comparando a ANOVA de mensurações repetidas, a ANOVA com escores de ganho, a ANOVA com *blocking* e a ANCOVA, esta última se mostrou como a mais indicada para mensuração dos resultados relativos à empregabilidade (resultado final). Para as variáveis aumento do número de oficinas e aumento do número de aprendizes (resultados intermediários), seria mais indicada a utilização da ANOVA com escores de ganho, uma vez que houve violação do pressuposto de independência entre o tratamento e o covariante e a ANOVA com *blocking* não apontou resultados satisfatórios. Portanto, a presente pesquisa evidencia alguns cuidados que têm que ser tomados por pesquisadores na utilização dos diferentes tipos de ANOVA, bem como da ANCOVA, para que possam contribuir de fato para que as avaliações de programas apresentem resultados mais claros e efetivos e com maior grau de certeza.

Referências

Antunes, C. M. F. (2001). Métodos epidemiológicos para planejamento e avaliação de impacto dos programas de controle vetorial da doença de Chagas. *Grupo de Trabajo OPS en Enfermedad de Chagas, Montevideo, Uruguay, Noviembre de 2001, 29-33*. Retirado em 12/11/2008 do sítio http://www.amro.who.int/spanish/ad/dpc/cd/consulta-6.pdf

Barreira, M. C. R. N. (2002). *Avaliação participativa de programas sociais*. São Paulo: Veras.

Brace, N., Kemp, R. & Snelgar, R. (2003). *SPSS for psychologists: A guide to data analysis using SPSS for Windows* (2ª. ed.). Hampshire: Palgrave Macmillan.

Bugarin, B. J. (2002). *Avaliação de programas públicos orientada para resultados: o papel dos órgãos de controle externo no Brasil*. Congresso sobre Reforma del Estado y de la Administración Pública promovido pelo Centro Latinoamericano de Administración para el Desarrollo – CLAD. Retirado em 02/03/2003 do sítio www.clad.org.ve/anales.

Campelo, A. F. (2006). Avaliação de programas sociais em ONGs: discutindo aspectos conceituais e levantando algumas orientações metodológicas sobre avaliação de impacto. *Interfaces de Saberes, 6,* 1-14.

Cano, I. (2002). *Introdução à avaliação de programas sociais.* Rio de Janeiro: FGV.

Carmines, E. J. & Zeller, R. A. (1979). *Reliability and validity assessments.* London: Sage.

Carvalho, S. N. (2003) Avaliação de programas sociais: balanço das experiências e contribuição para o debate. *São Paulo em Perspectiva*, *17*, 185-197.

Chianca, T. & Youker, B. (2004). La evaluacion en America Latina y el Caribe: Visión general de los desarrollos recientes. *Journal of Multidisciplinary Evaluation*, *1*, 89-102.

Cronbach, L. J. (1983). *Designing evaluations of educational and social programs.* Washington: Jossey-Bass.

Dancey, C. P. & Reidy, J. (2006). *Estatística sem matemática para Psicologia usando SPSS para Windows* (3ª. ed.) (L. Viali, Trad.). São Paulo: Artmed. (Trabalho original publicado em 1999)

Dugard, P. & Todman, J. (1995). Analysis of pre-test-post-test control group designs in educational research. *Educational Psychology*, *15*, 181-198.

Hedler, H. C. (2007). *Meta-avaliação das auditorias de natureza operacional do Tribunal de Contas da União: um estudo sobre auditorias de programas sociais*. Tese de Doutorado, Universidade de Brasília, Brasília.

Murta, S. G., Laros, J. A. & Tróccoli, B. T. (2005). Manejo de estresse ocupacional na perspectiva da área de avaliação de programas. *Estudos de Psicologia (Natal), 10*, 167-176.

Oskamp, S. (1981). *Applied Social Psychology.* Washington, D.C.: Prentice-Hall.

Posavac, E. J. & Carey, R. G. (2003). *Program Evaluation: Methods and case studies*. New Jersey: Prentice-Hall. Upper Saddle River.

Rossi, P. & Freeman, H. (1989). *Evaluácion: un enfoque sistemático.* New York: Sage.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Stevens, J. P. (2001). *Applied multivariate statistics for the social sciences* (4ª. ed.). Hillsdale, NJ: Erlbaum.

Sulbrandt, J. (1993). La evaluación de los programas sociales: una perspectiva crítica de los modelos usuales. *Centro Latinoamericano de Administración para el Desarrollo-CLAD (Caracas, Venezuela)*, 309-350.

Uchimura, K. Y., Bosi, M. L. M. (2002). Qualidade e subjetividade na avaliação de programas e serviços em saúde. *Caderno Saúde Pública, 18*, 1561-1569.

Wholey, J. S., Hatry, H. P. & Newcomer, K. E. (2004). *Handbook of practical program evaluation* (2ª. ed.). San Francisco, CA: Jossey-Bass.

Worthen, B. R., Sanders, J. R. & Fitzpatrick, J. L. (2004). *Avaliação de Programas: Concepções e práticas* (D. A. Azevedo, Trad.). São Paulo: Gente. (Trabalho original publicado em 1997)

Recebido em 09.04.2007
Primeira decisão editorial em 05.05.2008
Versão final em 13.06.2008
Aceito em 17.11.2008 ■