ESTUDOS EM AVALIAÇÃO EDUCACIONAL

ISSN **0103-6831**

# CONTRIBUIÇÕES DE HERALDO VIANNA PARA A AVALIAÇÃO EDUCACIONAL

FUNDAÇÃO CARLOS CHAGAS **V.25 N.60 NÚMERO ESPECIAL | DEZEMBRO 2014**

**ESTUDOS EM AVALIAÇÃO EDUCACIONAL • N. 2 JUL./DEZ. 1990 • FUNDAÇÃO CARLOS CHAGAS • SÃO PAULO**
Semestral
A partir de 2006 passa a ser QUADRIMESTRAL

ISSN 0103-6831
e-ISSN 1984-932X

1. Avaliação 2. Políticas Educacionais 3. Qualidade do Ensino
I. Fundação Carlos Chagas II. Departamento de Pesquisas Educacionais/FCC

**INDEXADO EM**

BAMP – Biblioteca Ana Maria Poppovic (*Brasil, FCC*)
*www.fcc.org.br*

BBE – Bibliografia Brasileira de Educação *(Brasil, Cibec/Inep/MEC)*
*http://pergamum.inep.gov.br/pergamum/biblioteca/index.php?resolution2=1024_1*

Clase – Citas Latinoamericanas en Ciencias Sociales y Humanidades *(México, Unam)*
*http://biblat.unam.mx/*

Edubase – *Faculdade de Educação (Brasil, Unicamp)*
*http://143.106.58.49/fae/default.htm*

Educ@ – Publicações on-line de Educação (*Brasil, FCC*)
*http://educa.fcc.org.br/scielo.php*

e-Revistas – Plataforma Open Access de Revistas Científicas Electrónicas Españolas y Latinoamericana (*Espanha*)
*http://www.erevistas.csic.es/*

Google Scholar
*http://scholar.google.com.br/*

Iresie – Índice de Revistas de Educación Superior e Investigación Educativa (*México, Cesu-Unam*)
*http://www.iisue.unam.mx/iresie/*

Latindex – Sistema Regional de Información en Línea para Revistas Científicas de América Latina, en Caribe, España y Portugal (*México, Unam*)
*http://www.latindex.unam.mx*

**VERSÃO ELETRÔNICA**
http://publicacoes.fcc.org.br

**VERSÃO IMPRESSA**
Dezembro de 2014
Tiragem: 300 exemplares

**E-MAILS**
eae@fcc.org.br *(contato)*
publicacoesfcc@fcc.org.br *(aquisição e assinaturas)*

**ESTUDOS EM AVALIAÇÃO EDUCACIONAL**

Periódico da Fundação Carlos Chagas criado em 1990 sucedendo *Educação e Seleção* (1980-1989). Publica trabalhos originais relacionados à educação, com perspectiva avaliativa, apresentados sob forma de relatos de pesquisa, ensaios teóricos, revisões críticas, artigos e resenhas.

As normas para a publicação de artigos e resenhas estão no final do volume.
A revista não se responsabiliza pelos conceitos emitidos em matérias assinadas.




**COORDENAÇÃO DE EDIÇÕES**
Adélia Maria Mariano da Silva Ferreira

**ASSISTENTE DE EDIÇÕES**
Camila Maria Camargo de Oliveira

**SECRETÁRIA DE EDIÇÕES**
Camila de Castro Costa

**PADRONIZAÇÃO BIBLIOGRÁFICA**
Biblioteca Ana Maria Poppovic

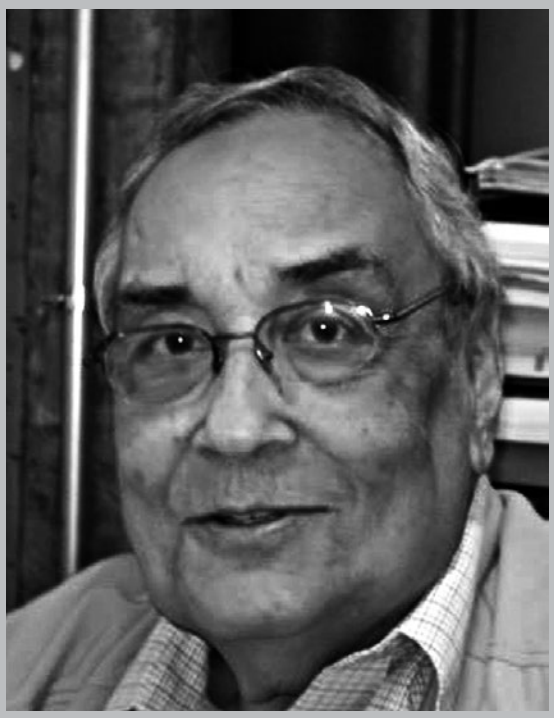
**REVISÃO ESTATÍSTICA**
Miriam Bizzocchi
Raquel da Cunha Valle

**PROJETO GRÁFICO E DIAGRAMAÇÃO**
Gustavo Piqueira | Casa Rex

**IMPRESSÃO**
Forma Certa Soluções Gráficas Personalizadas

# SUMÁRIO

## AS CONTRIBUIÇÕES DE HERALDO VIANNA PARA A AVALIAÇÃO EDUCACIONAL



Heraldo M. Vianna,
acervo da Fundação Carlos Chagas

# APRESENTAÇÃO

A revista *Estudos em Avaliação Educacional*, em seu número 60, apresenta uma homenagem ao seu idealizador, Heraldo Marelim Vianna, educador e pesquisador que desenvolveu sua carreira na Fundação Carlos Chagas. Editor deste periódico desde sua criação, em 1990, até sua aposentadoria, em 2008, foi um dos autores brasileiros que mais se dedicaram ao tema da avaliação educacional, quando esse ainda era muito incipiente no país. Foi, também, criador e editor da revista *Educação e Seleção*, igualmente publicada pela Fundação Carlos Chagas, entre 1980 e 1989.

Com graduação em Geografia e História, pela Universidade Estadual do Rio de Janeiro (UERJ), e em Direito, pela Universidade Federal do Rio de Janeiro (UFRJ), fez especializações nos Estados Unidos e na França, no campo da Educação, pela *University of Michigan Ann Arbor*, e em Pedagogia, pelo *Centre International D'études Pédagogiques*. Mestre e Doutor em Educação, pela *Michigan State University* e Pontifícia Universidade Católica de São Paulo, respectivamente, atuou como pesquisador no Departamento de Pesquisas Educacionais da Fundação Carlos Chagas de 1970 a 2007, tendo exercido cargos de gestão nesse departamento e no setor de testes e concursos. Foi membro do Conselho Estadual

de Educação de São Paulo, no período de 1997 a 2000, e prestou assessoria para o desenvolvimento de importantes avaliações de sistemas educacionais, tais como a dos estados de Minas Gerais e Paraná, e para o Exame Nacional de Cursos, por exemplo. A pedido do Ministério da Educação, desenvolveu, no final dos anos 1980, avaliações de rendimento educacional com alunos de escolas privadas e públicas envolvendo um número considerável de estudantes em diferentes regiões do país, tarefa quase hercúlea naqueles tempos em que as informações educacionais eram desatualizadas e pouco robustas. Foram trabalhos que permitiram antever importantes políticas de avaliação, como o Sistema de Avaliação da Educação Básica (Saeb), por exemplo. Foi responsável, também, pelo desenvolvimento de avaliações inovadoras, como *The International Assessment of Educational Progress* (IEAP), avaliação em ciências e matemática envolvendo alunos de vinte países, e Avaliação de habilidades de vida.

Sua obra envolve mais de 60 artigos e 15 livros sobre educação, sendo fortemente voltada para a avaliação educacional. Sua preocupação constante era oferecer aos educadores brasileiros elementos teóricos, contextualizados historicamente, para a compreensão do desenvolvimento desse campo de estudos. Muitos de seus artigos iniciais são voltados para aspectos psicométricos, ou seja, sobre como elaborar e tratar os resultados de aplicações de instrumentos de avaliação. Com a criação de *Estudos em Avaliação Educacional*, passou a escrever artigos mais centrados na disseminação das teorias e modelos que embasam as avaliações de programas e sistemas educacionais. Alguns textos visaram a difundir resultados de avaliações realizadas no Brasil, outros ofereceram reflexões pertinentes e atuais sobre os impasses e problemas que envolvem a avaliação educacional.

Neste ano de 2014, a Fundação Carlos Chagas completa 50 anos de existência e, naturalmente, almeja homenagear aqueles que lutaram pela construção de sua sólida reputação de instituição inovadora, competente e séria. Não poderia deixar de resgatar a contribuição de Heraldo Marelim Vianna. Surgiu, assim, a ideia de reeditar alguns de seus textos. Para a composição de *Estudos em Avaliação Educacional* nº 60 foram selecionados 14 artigos do autor, publicados em periódicos da Fundação Carlos Chagas, especificamente *Cadernos de Pesquisa*, *Estudos em Avaliação Educacional*, *Educação e Seleção* e *Textos FCC*. Os artigos

estão separados por temas: teoria e história da avaliação educacional; formação do avaliador; instrumentos de avaliação; e reflexões sobre a prática avaliativa.

Em **Avaliação educacional: teoria e história** são apresentados quatro textos. O primeiro, "Avaliação educacional: uma perspectiva histórica", de 1995, discorre sobre o desenvolvimento da avaliação educacional nos Estados Unidos e na Inglaterra, e, posteriormente, no Brasil. O texto apresenta autores e conceitos importantes para interessados na área. A seguir, em "Medida da qualidade em educação: apresentação de um modelo", de 1990, Heraldo Vianna discute o conceito de qualidade em educação, suas implicações e relações com a responsabilização educacional. No terceiro artigo deste segmento, "Avaliação de programas educacionais: duas questões", de 2005, o autor discute aspectos da validade em avaliação de programas, apontando as diferenças entre esse tipo de avaliação e a avaliação de desempenho, destacando o aspecto democrático desse processo e ressaltando a importância da disseminação dos resultados para seu efetivo impacto. Em "Fundamentos de um programa de avaliação educacional", de 2003, o autor reflete sobre aspectos que envolvem a definição de uma política de avaliação do sistema educacional brasileiro, considerando a diversidade socioeconômica e cultural dos alunos, os problemas para a disseminação dos resultados e as necessidades envolvidas no planejamento escolar e na tomada de decisões.

A segunda parte, **Avaliação educacional: formação do avaliador**, é composta por apenas dois textos, mas que traduzem as condições mais relevantes a serem consideradas quando o foco é quem desenvolve a avaliação. Em "Avaliação educacional: problemas gerais e formação do avaliador", de 1982, Vianna discorre sobre a complexidade crescente do campo da avaliação educacional e suas consequências para a formação do avaliador. Dentre os aspectos discutidos estão os modelos teóricos desenvolvidos por alguns autores, as diferenças e semelhanças entre pesquisa e avaliação educacional e as funções do avaliador educacional. Em "Avaliação e o avaliador educacional: depoimento", de 1999, Heraldo relata as experiências vivenciadas por ele no campo avaliativo e as reflexões que suscitaram, sobretudo quanto às necessidades na sua formação e na revisão de seus posicionamentos.

Um terceiro grupo de textos refere-se a **Instrumentos de avaliação educacional**. Em "Qualificação técnica e construção de instrumentos de medida educacional", de 1984, Vianna trata dos problemas técnicos que frequentemente são observados entre instrumentos de medida do rendimento escolar utilizados no meio educacional. Para solucionar o problema, sugere modelos de análise de medidas e tópicos para formação de educadores que abordem o tema. Em "Natureza das medidas educacionais", de 1984, o autor trata das dificuldades inerentes à mensuração de variáveis educacionais, das medidas possíveis (atributos, efeitos) e das diferentes formas de medi-las (escalas nominais, ordinais, intervalares e de razão), bem como reflete sobre as divergências quanto ao significado e interpretação das medidas. O texto "Validade de construto em testes educacionais", de 1983, é de interesse muito atual, tendo em vista que muitos testes educacionais visam a medir processos mentais complexos. O artigo apresenta metodologias diversificadas para validação de testes e de teorias, diferenciando os conceitos de fidedignidade e validade. Há, também, um texto sobre "Aplicação de critérios de correção em provas de redação", de 1978, que discorre sobre método de validação do treinamento de professores para correção desse tipo de prova em vestibulares.

Finalmente, uma quarta parte trata de **Reflexões sobre a prática avaliativa**. Em "Avaliando a avaliação: da prática à pesquisa", de 1992, Vianna faz uma retrospectiva do desenvolvimento da avaliação no Brasil e destaca a importância de se criar uma cultura da avaliação aliada à pesquisa, de forma a lhe dar credibilidade. Em seguida, o artigo "A prática da avaliação educacional: algumas colocações metodológicas", de 1989, discute as diferenças entre medir e avaliar, destacando que medir com fidedignidade não significa realizar uma boa avaliação, pois é preciso haver precisão e validade de diferentes tipos. Em "Avaliações nacionais em larga escala: análises e propostas", de 2003, Heraldo analisa avaliações que abrangem grande número de pessoas, tais como vestibulares e as avaliações sistêmicas desencadeadas no Brasil a partir dos anos 1990, e constrói um interessante panorama do desenvolvimento da cultura avaliativa no país. O último texto selecionado para este número especial de *Estudos em Avaliação Educacional* é "Avaliação educacional: vivência e reflexão", publicado em 1998. Bastante crítico, Vianna questiona

o uso de metodologias sofisticadas de análise que dificultam o entendimento dos resultados pelos maiores interessados – professores e alunos –, ressalta a necessidade do uso conjunto de técnicas qualitativas e quantitativas que apresentem precisão e validade nos seus métodos e resultados, bem como ressalta a responsabilidade pública do avaliador na elaboração de modelos avaliativos robustos e factíveis, que respondam às necessidades dos atores educacionais envolvidos.

A obra de Heraldo Marelim Vianna é vasta e pode gerar outras publicações temáticas. Com este número ensejamos convidar o leitor a conhecer e refletir sobre os pressupostos teóricos e metodológicos que envolvem a avaliação educacional, as necessidades de formação e de revisão das práticas avaliativas, a fim de contribuir para o debate frente à maior complexidade e aos desafios atuais da realidade educacional.

Muito devo a esse pesquisador, com quem trabalhei desde que me formei bacharel em Psicologia, na Pontifícia Universidade Católica de São Paulo (PUC-SP). Como bolsista nos projetos de avaliação, com ele aprendi, primeiramente, a gostar dos números e das técnicas estatísticas e psicométricas; depois, ao nos depararmos com a complexidade da educação brasileira, foi inevitável a incorporação de técnicas mais qualitativas, que permitissem captar sutilezas e dar aprofundamento a hipóteses e pistas. Como assistente de pesquisa e pesquisadora do Departamento de Pesquisas, participei com o Prof. Heraldo de programas de avaliação de diferentes portes, enfoques e finalidades. Sinto-me honrada em homenageá-lo como organizadora deste número especial de *Estudos em Avaliação Educacional*, esperando contribuir para a formação de novas gerações de avaliadores, que prezem o rigor e a qualidade como sempre o fez esse educador.

*Gláucia Torres Franco Novaes*
*Editora de* Estudos em Avaliação Educacional

# AVALIAÇÃO EDUCACIONAL: TEORIA E HISTÓRIA

# AVALIAÇÃO EDUCACIONAL: UMA PERSPECTIVA HISTÓRICA[1]

## 1. INTRODUÇÃO

A pesquisa e a avaliação têm um significado especial no delineamento do processo decisório em educação. A pesquisa, inicialmente, e, depois, a avaliação, especialmente a avaliação de programas, projetos e produtos, passaram a ter uma dimensão maior a partir do século XX, sofrendo ambas – pesquisa e avaliação – a influência de diferentes ciências, como por exemplo, a psicologia, a psicometria, a sociologia, a antropologia, a etnografia e a economia, entre outras, de determinaram novos enfoques metodológicos com base em vários posicionamentos teóricos.

A avaliação, após o trauma provocado pela constatação da deficiência tecnológica associada à carência educacional, no mundo ocidental, especialmente nos Estados Unidos, com o lançamento do Sputnik, no dia 4 de outubro de 1957, tornou-se impositiva. Há todo um esforço no sentido de recuperar o tempo educacional que fora perdido, sendo criados novos currículos, e a avaliação passou a ter papel de relevância no desenvolvimento de novas estratégias de ensino[2].

A literatura sobre avaliação educacional, a partir dos anos 60, cresce enormemente. Surgem novos posicionamentos

**1** Artigo publicado na revista *Estudos em Avaliação Educacional*, São Paulo, n. 12, p. 7-24, jul./dez. 1995.

**2** Alguns desses novos currículos em Matemática (*Mathematics Study Group* – MSG), em Física (*Physical Science Study Committee* – PSSC), em Química (*Chemical Bond Approach* –CBA) e em Biologia (*Biological Study Committee* – BSC), entre outros, foram traduzidos e adaptados ao contexto nacional pela Fundação Brasileira para o Ensino de Ciências (FUNBEC) e no Centro de Treinamento de Professores de Ciências (CECISP), graças aos esforços dos professores Isaías Raw, Myrian Krasilchik e Norma Maria Cleffi, com a colaboração de professores brasileiros, no início da década de 60.

teóricos e novas propostas de atividades práticas, que enriquecem e tornam complexo o campo da avaliação. Algumas questões fundamentais são levantadas, com profundas implicações na sua teoria e na sua prática. É justo, portanto, que, inicialmente, seja discutida a sua evolução histórica a partir da perspectiva norte-americana, em cujo contexto se desenvolveram numerosos e importantes trabalhos, em particular em decorrência da ação seminal de Ralph W. Tyler. Outras fontes, entretanto, também serão indicadas, especialmente inglesas e, subsidiariamente, trabalhos de autores nacionais, ainda que a produção teórica destes últimos seja bastante reduzida no seu aspecto quantitativo.

### 1.1. AVALIAÇÃO EDUCACIONAL NOS ESTADOS UNIDOS: UM ESBOÇO

A sociedade norte-americana sofreu grandes transformações estruturais em decorrência do impacto advindo da Revolução Industrial e, a exemplo do que ocorreu na Inglaterra, o poder público passou a discutir diferentes programas sociais, inclusive de natureza educacional. A avaliação nos Estados Unidos possui uma tradição de quase dois séculos, ainda que o seu momento mais intenso tenha ocorrido a partir da década de 1960. É também nesse período que a avaliação começa a se definir como uma profissão estruturada. A fim de chegar a essa situação, a avaliação atravessou diferentes momentos, conforme Madaus, Stufflebeam e Scriven (1993).

A avaliação associada ao processo educativo sempre existiu sob diferentes formas (WORTHEN; SANDERS, 1987), mas somente adquire uma natureza formal quando Horace Mann (1845) inicia a prática da coleta de dados para a fundamentação de decisões de políticas públicas que afetam a educação. A avaliação, aos poucos, começa a desenvolver procedimentos ainda utilizados nos dias de hoje, como o *survey*, e instrumentos objetivos padronizados. Além disso, determinadas práticas ainda persistem, porque baseadas na obtenção de escores dos alunos como principal elemento para a avaliação da eficiência de programas educacionais, aliás, traço característico das avaliações realizadas no atual contexto brasileiro (1995).

Houve, nessa época, como mostram Madaus *et al.* (1993), uma politização dos dados obtidos através de provas escritas, que foram usados para comparar escolas e promover o afastamento de diretores que se opunham ao programa de eliminação de castigos físicos que então existiam, e cuja supressão era defendida por Samuel G. Howe e Horace Mann. A primeira avaliação propriamente dita somente teve lugar bem mais tarde, no final do século XIX, entre 1887 e 1898, quando Joseph Rice procurou verificar a influência do tempo dedicado a exercícios (*drill*) no processo de alfabetização (*spelling*) em diferentes unidades escolares. Ainda que esse trabalho de Rice tenha tido uma grande repercussão na época e levado muitos professores a mudarem seus procedimentos de alfabetização, na verdade, o maior impacto desta avaliação estava no seu caráter experimentalista e quantitativo, procedimento até então inédito ainda segundo o destaque de Madaus *et al.* (1993). Rice, de certa forma, antecipou-se a ideias sobre pesquisa experimental em educação que somente seriam defendidas na segunda metade do século XX, a partir dos anos 50, por Lindquist (1953) e Campbell e Stanley (1963).

Os trinta primeiros anos da vida social norte-americana no século XX sofreram a influência de três elementos desenvolvidos em princípio para o gerenciamento industrial: sistematização, padronização e eficiência, que acabaram por afetar a totalidade da sociedade, inclusive na área educacional. A comunidade educacional como um todo passa, então a preocupar-se com o desenvolvimento de uma metodologia que permitisse medir a eficiência dos seus professores, a construir instrumentos e a definir padrões que possibilitassem a mensuração do grau de eficiência das suas escolas e dos diversos sistemas educacionais, seguindo, assim, no campo da educação, os procedimentos que empresários procuravam implantar no mundo da indústria. Aproximadamente quase cem anos depois, a educação brasileira, muitas vezes por influência de agências financiadoras externas, começa a se preocupar com os mesmos problemas ligados diretamente ao processo de gerenciamento das instituições educacionais, conforme poder-se-á observar no item final do presente trabalho.

O grande passo na evolução da avaliação educacional foi dado por E. L. Thorndike, nos princípios do século XX, ao desenvolver toda uma fundamentação teórica sobre a possiblidade de

medir mudanças nos seres humanos (WORTHEN; SANDERS, 1987). A partir desse momento, todo um aparato tecnológico é desenvolvido para a medida de capacidades humanas e a avaliação, em consequência, passa a ter o significado de medida (*testing*), concepção que ainda prevalece em amplos setores educacionais, inclusive fora dos Estados Unidos, como no Brasil.

Madaus, Stufflebeam e Scriven (1993) mostram claramente que essa é a época do *survey*, que emprega diferentes critérios (taxa de evasão, taxa de aprovação) para medir a eficiência da escola e/ou professor. O uso do *survey* leva, naturalmente, ao desenvolvimento de diversos tipos de **testes objetivos** nas várias áreas curriculares. Esses testes, por sua vez, eram **referenciados a objetivos** e deram origem aos **testes referenciados a normas**, sendo, nesse caso, a porcentagem de alunos aprovados nos testes o critério adotado para julgar a eficiência da escola/professor. Além disso, é nesta mesma época, há quase um século, que começam a surgir testes normativos, elaborados por importantes cientistas educacionais, como foi o caso de Edward Thorndike, com vistas ao desenvolvimento de um instrumental que possibilitasse a comparação entre sistemas. Os **testes padronizados**, bastante característicos da cultura educacional norte-americana, mas que, felizmente, ainda não chegaram ao meio educacional brasileiro, surgiram logo após o conflito mundial de 1914/18 e passaram a ser usados, equivocadamente, na determinação da eficiência de programas educacionais e na apresentação de diagnósticos relativos a currículos e sistemas educacionais, além de servirem, também, para a tomada de decisões sobre o desempenho escolar dos alunos.

Avaliação e medida por meio de testes confundem-se nessa época – e ainda continuam a se confundir nos dias atuais-; por outro lado, começam a surgir instituições especializadas na realização de *surveys* na área educacional. Estas instituições constituíram o embrião que geraria futuros e importantes centros educacionais, criados nos anos 60 e 70, nas universidades, com o objetivo de realizaram trabalhos avaliativos em diferentes áreas, inclusive na educacional. Os *surveys*, inicialmente, eram locais, limitados aos *school districts*, ainda segundo Madaus *et al.* (1993). No decorrer dos anos 50 e 60 é que começam a surgir efetivamente nos Estados Unidos os primeiros estudos de currículo a nível

nacional. Os problemas passam a ser considerados não apenas em função do interesse local, mas levando em conta o contexto nacional, com uma audiência bem mais ampla do que aquela dos *school districts*. E, assim, por via de consequência, complexos problemas ligados à epistemologia da pesquisa/avaliação educacional começam a ser discutidos por educadores e avaliadores, que, entre outros aspectos, passam a se preocupar com a questão da **generalizabilidade** das suas conclusões, como mostram Madaus *et al.* (1993) e Norris (1993).

A influência de Ralph W. Tyler durante o período de 1930 a 1945 foi considerável e, assim, com justa razão, passou a ser considerado o verdadeiro iniciador da avaliação educacional. A sua ação foi bastante ampla, influindo na educação em geral, especialmente em assuntos ligados à teoria, à construção e à implementação de currículos, que procurou conceituar como um conjunto de experiências educacionais diversificadas que deveriam ser planejadas de forma a levar os alunos à concretização de determinados objetivos. A avaliação educacional, cujo termo foi por ele criado, objetivaria que professores aprimorassem seus cursos e os instrumentos de medida que construíssem pudessem verificar a congruência entre os conteúdos curriculares e as capacidades desenvolvidas.

A sociedade norte-americana depois do término da Segunda Guerra Mundial (1939-45) atravessou um período bastante crítico, envolvendo problemas econômicos e sociais (racismo/segregacionismo). Houve, por outro lado, como decorrência da chamada "guerra fria", resultante do conflito ideológico-maniqueísta entre o capitalismo ocidental e o socialismo real do leste europeu, a criação de um grande complexo militar-industrial que atuaria intensamente na vida americana por largo tempo até o início dos anos 90, com a crise que levaria à desagregação da União Soviética. Apesar da crise socioeconômica, especialmente no período inicial, entre 1946 e 1957, importantes transformações ocorreram na educação americana.

O problema da avaliação educacional foi bastante discutido, somas consideráveis de dados foram levantadas, mas quase não foram utilizadas para a solução de problemas, como ressaltaram Madaus *et al.* (1993). É nesse período que ocorre o desenvolvimento da teoria clássica dos testes e o surgimento de novas abordagens sobre avaliação educacional, como uma rea-

ção ao modelo proposto por Tyler. É importante destacar, nesse período, o aparecimento de uma instituição atuante até os dias fluentes e que serviu de modelo para a criação de outros órgãos semelhantes nos seus objetivos: – o ***Educational Testing Service*** **–** ETS (1947), por inspiração de vários educadores, destacando-se E. F. Lindquist e Ralph W. Tyler. O **ETS**, a partir de sua criação, passaria a ter influência decisiva no desenvolvimento de testes (padronizados) e em amplos programas de avaliação, como o ***National Assessment of Educational Progress*** (NAEP) e o ***International Assessment of Educational Progress*** (IAEP). A proliferação dos testes, nem sempre construídos segundo princípios claros e bem definidos cientificamente, levou algumas instituições representativas (*American Psychological Association* – APA, *American Educational Research Association* – AERA, *National Committee on Measurement in Education* – NCME, sobretudo) a estabelecerem recomendações para a construção de testes psicológicos e técnicas de diagnóstico (1954) e para a elaboração de testes de rendimento escolar (1955), – o que nem sempre é seguido, inclusive no Brasil, que desconhece essas recomendações, salvo as exceções de sempre; mais tarde, após sucessivas revisões, essas mesmas recomendações se transformaram em padrões (*standards*) a serem seguidos na área de construção dos vários instrumentos psicométricos.

A avaliação educacional nos princípios dos anos 60, nos Estados Unidos, começa a incidir sobre grandes projetos de currículos financiados com o apoio federal, surgindo, nesse momento histórico, a figura do avaliador como um profissional com atividades específicas até então exercidas por educadores com formação generalista. É o início de uma época de especialização – que ainda não chegou ao contexto brasileiro – em que as universidades começam a se preocupar com a formação específica de recursos humanos efetivamente qualificados para os trabalhos de avaliação.

Apesar do envolvimento das melhores cabeças pensantes na área educacional e dos grandes fundos alocados a múltiplos projetos, os resultados das avaliações durante a década de 60, ainda que realizados com grande rigor técnico, não se revelaram plenamente satisfatórios e surgiram diversas reações. É nesse contexto que o ***Teachers College Record*** publica importante artigo de Cronbach (1963) criticando a situação e pro-

pondo novos caminhos para a prática da avaliação, conforme veremos. Apesar da advertência de Cronbach de que a avaliação da eficácia de um currículo se deveria processar ao longo de sua estruturação e não depois de sua conclusão, o impacto das suas ideias não foi imediato, mas levou especialistas a discutirem detidamente seus problemas profissionais, como mostraram Worthen e Sanders (1987).

O trabalho de Cronbach – ***Course Improvement through Evaluation*** (1963) – levantou importantes questões metodológicas, inclusive a relacionada com a análise direta dos resultados dos itens em substituição à análise concentrada em escores globais, além de abordar outros problemas de natureza conceitual que continuam válidos na atualidade e mostrar as limitações das avaliações *post-hoc* no desenvolvimento de currículos, conforme foi referido anteriormente. Esse artigo mostra os fundamentos básicos da prática da avaliação e abre novas e estimulantes perspectivas para o trabalho do avaliador no seu dia a dia como profissional.

O contexto político decorrente, inicialmente, da administração John Kennedy e, a seguir, após 1963, da gestão Lyndon Johnson, em que houve uma grande preocupação com os reflexos das desigualdades sociais na diferenciação das oportunidades educacionais, passa a determinar um grande esforço no sentido de criar novas condições na área educacional, surgindo, por influência de Robert Kennedy, o conceito de responsabilidade em educação (*accountability*), a fim de evitar possíveis desperdícios dos recursos financeiros concedidos aos programas curriculares e a suas avaliações, na área da educação compensatória. A avaliação deixa, nesse novo contexto, de ser apenas um trabalho teórico de alguns educadores, transformando-se numa prática constante, que, em muitos casos, assume um caráter quase obsessivo na cultura educacional norte-americana.

A aprovação de importante lei sobre educação elementar e secundária – ***Elementary and Secondary Education Act (ESEA)***, em 1965, vai gerar um novo quadro que favorecerá a pesquisa e a avaliação educacional, tendo em vista a imposição legal de que todos os projetos financiados deveriam ser obrigatoriamente avaliados (WORTHEN; SANDERS, 1987), o que obrigou os educadores a dedicarem grandes esforços para a avaliação de suas atividades e a elaboração de relatórios passou a ser uma

prática habitual, no controle da qualidade do ensino e dos investimentos feitos em educação.

A situação criada após 1965 mostrou que os educadores norte-americanos não estavam preparados para os grandes desafios da avaliação, como, aliás, ainda não estão suficientemente preparados os educadores brasileiros para o enfrentamento dos grandes problemas da avaliação de hoje (1995), havendo muito teorização abstrata sobre o assunto, mas pouca realização prática. Professores norte-americanos, sob o impacto do **ESEA** em 1965, improvisaram-se em avaliadores. Worthen e Sanders (1987) mostraram que mesmo aqueles que tinham alguma *expertise* técnica não tinham treinamento suficiente em planejamento, medidas e estatísticas, quadro que, no momento, ocorre no Brasil. Isso produziu avaliações bastante comprometidas, que pouco serviram para avaliar a efetiva eficiência dos sistemas educacionais, seus currículos e programas.

A prática generalizada da avaliação mostrou que muitos instrumentos e estratégias utilizadas em avaliação não eram adequados aos propósitos definidos, ficando, inclusive, caracterizada a inadequação do emprego de testes padronizados na avaliação de programas, conforme a discussão de Madaus *et al.* (1993). Essa insatisfação acabou tendo efeitos positivos. Novas teorias sobre avaliação começam a despontar e todo um trabalho de reformulação da prática da avaliação é empreendido, destacando-se as figuras exponenciais de M. Scriven (1967), R. Stake (1967) e D. Stufflebeam (1971), cujos trabalhos vão dar uma nova dimensão metodológica à avaliação educacional. A avaliação de alguns programas de impacto, como o ***Head Start*** e o ***Sesame Street***, criados com objetivo de eliminar desequilíbrios educacionais decorrentes da origem social e racial, geram, na verdade, acentuadas controvérsias, conforme a colocação de Madaus, Airasian e Kellaghan (1980), sobretudo face ao baixo desempenho das crianças urbanas das áreas de periferia. A avaliação desses e de outros programas sociais teve um grande impacto social, ao apresentar resultados pouco promissores, e serviu para aprofundar as críticas a muitos procedimentos então em uso e possibilitou o desenvolvimento de novas metodologias.

A partir dos anos 70, a avaliação educacional torna-se um campo profissional definido, exigindo especialização aprofundada, com a exclusão de improvisações supostamente técnicas, mas

pouco sólidas conceitualmente, num campo em que atuavam diferentes profissionais, inclusive administradores. A avaliação, pelo menos no contexto norte-americano, deixa de ser uma "terra de ninguém". Simultaneamente, surgem importantes revistas especializadas, algumas de alta qualidade técnica, associando avaliação e políticas públicas, avaliação e planejamento, entre outras, permitindo, assim, a disseminação de novas ideias, a formulação de teorias e modelos, e sobretudo, a divulgação de importantes estudos.

A literatura sobre avaliação educacional – distinta da bibliografia sobre testes e medidas – torna-se copiosa e algumas universidades, como por exemplo, a de Illinois (Urbana), a de Stanford (Palo Alto, Califórnia) e a *Western Michigan University* (Kalamazoo, MI), entre outras, desenvolvem intensos programas de formação de avaliadores, ressaltando-se, ainda, o trabalho de Ben S. Bloom em Chicago e o de W. James Popham, na *University of Califórnia* (UCLA). Além disso, numerosos centros de avaliação foram disseminados por todo o território norte-americano, conforme o longo elenco de instituições apresentado por Madaus *et al.* (1993).

Apesar dos avanços e de todo aparato técnico desenvolvido, a área da avaliação ainda não está exaurida em todas as suas potencialidades, novas perspectivas de ação estão sendo abertas e a radicalização positivismo/quantitativo em relação ao fenomenológico/qualitativo é menos radical, sendo as divergências porventura ainda subsistentes na verdade um reflexo de diferenças ideológicas e menos conflitos metodológicos, conforme ressaltam Madaus *et al.* (1993).

Finalmente, após a longa exposição sobre a avaliação na perspectiva norte-americana, queremos ressaltar que, hoje, nos Estados Unidos, há uma grande preocupação com a qualidade dos trabalhos de avaliação, tendo sido fixados padrões a serem seguidos conforme o posicionamento do ***Joint Committee on Standards for Educational Evaluation***, coordenado por D. Stufflebeam. Ainda que a avaliação educacional, suas teorias e modelos deem margem a grandes controvérsias, é, sem dúvida, um campo dinâmico, em constante transformação para atendimento das múltiplas exigências da qualidade em educação.

## 1.2. AVALIAÇÃO EDUCACIONAL NA INGLATERRA: ALGUMAS EXPERIÊNCIAS

A avaliação educacional na Inglaterra surge em meados do século XIX, integrando um programa social maior com vistas à eficiência nacional. A educação, na concepção inglesa, é considerada como elemento indiscutivelmente essencial para alcançar essa eficiência. As pesquisas (e a avaliação) na área educacional, inicialmente, desenvolvem uma linha ligada a métodos estatísticos e à técnica de *survey*, como ocorreu também nos Estados Unidos, contribuindo, assim, para o desenvolvimento de uma tecnologia: – a psicometria, que mais tarde, seria duramente criticada por muitos educadores ingleses.

Os trabalhos realizados por F. Galton, K. Pearson, C. Spearman e C. Burt, entre outros, na Inglaterra, contribuíram para que a psicometria tivesse influência considerável na avaliação educacional, em especial na construção de instrumentos de medidas psicológicas e do rendimento escolar. A utilização da "teoria dos erros" e a aplicação da noção de "distribuição normal" no estudo da variabilidade das diferenças individuais concorreu para gerar reações justificadas ao uso da psicometria da avaliação[3]. Isso contribuiu para que novos caminhos se abrissem para a exploração de diversos aspectos da avaliação educacional. Entretanto, apesar dessas reações adversas à tradição psicométrica, a Inglaterra, a partir da década de 50, realizou importantes trabalhos sobre a alfabetização e a aprendizagem da leitura, contribuindo, desse modo, para o delineamento de novas metodologias experimentais e o desenvolvimento da teoria dos testes (NORRIS, 1993).

**3** Para um estudo detalhado da influência da Psicometria na evolução das medidas educacionais ver J. Rust e S. Golombok. *Modern Psychometrics – The Science of Psychological Assessment*. London: Routledge, 1992.

A teoria do desenvolvimento econômico, pleno emprego e segurança nacional, inspirada em concepções apresentadas por Keynes, teve amplas repercussões na definição das políticas públicas da educação inglesa. Houve, naquele momento histórico, uma preocupação especial com os problemas relativos à qualidade dos estudantes que, em grande número, procuravam cursos de ciências aplicadas e tecnologias, e, concomitantemente, um profundo interesse na formação de professores de ciências destinados às escolas secundárias. A partir do ideário econômico de Keynes, que em grande parte fundamenta o capitalismo posterior à guerra de 1939-45, inclusive nos Estados Unidos, os educadores ingleses passaram a considerar a relação entre mão de obra tecnicamente qualificada pelo processo educacional e o crescimento

econômico/segurança nacional/e desenvolvimento tecnológico. A Inglaterra, por outro lado, constatou a inteira obsolescência dos currículos de suas escolas, o que tinha grande impacto nos programas que visavam à formação de mão de obra qualificada. A mudança desse quadro se impunha, a fim de proporcionar recursos humanos qualificados capazes de garantir o crescimento econômico. À educação caberia essa tarefa, ainda que as transformações no sistema educacional não pudessem ser imediatas, tendo em vista a impossibilidade de o governo inglês influir diretamente em uma estrutura altamente descentralizada.

Havia, desde meados da década de 50, uma consciência de que a Inglaterra não estava desenvolvendo capacidades e formando cientistas da mesma forma como outros países industrializados. A partir de 1962, graças à associação do Governo com a Fundação Nuffield, instituição de caráter privado, iniciaram-se atividades para a reformulação dos currículos e a formação de professores de ciências. A avaliação, no início dos anos 60, estava mais ligada à pesquisa educacional e os responsáveis pelo Projeto Nuffield revelaram uma certa relutância em aceitar a ideia de uma avaliação do Projeto, conforme Norris (1993); porquanto, argumentavam, o interesse maior estava na mudança de atitudes dos estudantes em relação à ciência e não na medida de possíveis ganhos de conhecimento fatual. A qualidade do Projeto, na visão da Fundação Nuffield, era bastante clara e para uma conclusão a respeito do valor desse empreendimento o Projeto dependeria, apenas, de **informações** das escolas que o utilizavam, das **críticas** que seriam oferecidas e do conhecimento dos **problemas** constatados durante sua execução.

Ainda que uma avaliação externa tenha sido afastada, em virtude da impossibilidade de estabelecer comparações entre o Projeto Nuffield e outros tipos de abordagem curricular, uma avaliação estava sendo feita na realidade, via informações, críticas e análise dos problemas, conforme foi apontado. Uma avaliação do tipo que se tornaria bastante utilizado por amplos segmentos educacionais ingleses: – a avaliação qualitativa. Ou seja, uma avaliação não-quantitativa, na tradição psicométrica, mas uma avaliação igualmente válida do ponto de vista científico. Isso ocorreu na parte de ciências do Projeto, na de línguas (francês).

A ***National Foudation for Educational Research*** (1964) reali-

zou uma avaliação baseada em testes e questionários, a fim de obter dados psicométricos e sociométricos, com estudos comparativos de vários grupos. Apesar da pouca repercussão dos resultados na comunidade educacional – o que ocorre com bastante frequência, inclusive, ou sobretudo, no Brasil –, ficou evidenciada a coexistência do quantitativo e do qualitativo no contexto educacional inglês, tendência que ainda hoje existe, apesar da falsa imagem divulgada por alguns de que haveria uma total dominância da pesquisa e da avaliação qualitativa, simplesmente.

A partir de 1966, ainda segundo o trabalho desenvolvido por Norris (1993), iniciaram-se estudos avaliativos sobre os Projetos Nuffield (ciências e matemática) por intermédio dos ***Schools Council***, agência autônoma fundada pelo governo central e governos locais. Algumas colocações do projeto avaliativo são extremamente importantes do ponto de vista metodológico. Assim, a avaliação era vista não apenas como uma medida dos resultados do projeto, mas como parte integrante do próprio projeto de construção do currículo, segundo a perspectiva apresentada por Cronbach (1963), A avaliação, por outro lado, em relação ao currículo não deveria ser vista como uma simples apresentação de meros exercícios para que um produto final fosse avaliado, mas, na verdade, a avaliação consistiria na coleta e no uso de informações que possibilitariam decisões sobre as diferentes fases do desenvolvimento de um programa educacional.

Ao destacarem as relações entre avaliação e tomada de decisões, as diretrizes estabelecidas para os ***Schools Council*** deixaram perfeitamente claro os momentos em que a avaliação educacional poderia emprestar sua contribuição: – 1) nas decisões sobre o conteúdo dos cursos e relativas aos métodos a empregar para o seu ensino; – 2) nas decisões sobre as necessidades dos alunos; e – 3) nas decisões concernentes ao treinamento de professores.

O projeto do ***Schools Council***, ainda segundo Norris (1993), destaca a participação dos avaliadores, muitas vezes apenas professores, e não necessariamente especialistas no campo da avaliação, desde o início dos trabalhos avaliativos e acentua a concepção de uma avaliação para tomada de decisões, aspecto que é expresso na maioria das diferentes abordagens sobre avaliação desenvolvidas na Inglaterra, e que merece especial destaque no modelo de Stufflebeam, nos Estados Unidos. A

avaliação, na visão do ***Schools Council***, é responsabilidade de um grupo de avaliadores, que nem sempre são avaliadores profissionais, e não de apenas um único indivíduo.

Avaliação educacional (***evaluation*** – avaliação de programa/ ***assessment*** – avaliação do rendimento escolar) é um assunto altamente controvertido na Inglaterra, apresentando tendências várias que quase sempre se opõem à avaliação somativa, apegando-se mais à do tipo individualizada. A avaliação também entra em choque com um dos mitos caros à educação inglesa – a autonomia dos professores. Observa-se, ainda, que várias abordagens têm um conteúdo sociológico, refletindo uma posição aceita por expressivo número de educadores. Os que seguem essa linha mostram a existência de uma correlação entre o socialmente carente e o rendimento escolar, advogando que os responsáveis pela educação devem apresentar uma discriminação positiva a favor daqueles que se acham nessa situação. Essa tese parece-nos de certa forma enviesada, pois, na realidade, a correlação não se limita ao carente, ela simplesmente existe, independentemente da condição social, tornando-se, entretanto, evidente entre os mais desfavorecidos economicamente[4].

**4** Uma apresentação detalhada da evolução histórica da avaliação na Inglaterra e suas mais recentes experiências de avaliação de programas acha-se em Norris (1993), capítulos 2, 4, 5 e 6.

### 1.3. AVALIAÇÃO EDUCACIONAL NO BRASIL (1960-95)

As atividades de avaliação educacional no Brasil são bastante escassas, ainda que, no momento atual (1995), a temática venha sendo verbalizada pelo Ministério da Educação, que tem apresentado ideias, às vezes bastante discutíveis, sobre a seleção para o acesso ao 3º grau de ensino e relativas à avaliação institucional por intermédio de uma avaliação do rendimento acadêmico ao término dos cursos de graduação.

A avaliação educacional, conforme será discutido, mesmo quando a nível de sistema, no contexto brasileiro, baseia-se, fundamentalmente, no rendimento escolar, ainda que haja uma coleta simultânea de dados socioeconômicos e de variáveis ligadas ao ensino, ao professor e à escola, em alguns casos. As pesquisas e estudos de avaliação apresentados nesse item em discussão não visam a uma revisão da literatura e nem incluem trabalhos acadêmicos, quase sempre relacionados a pré-requisitos da pós-graduação universitária, como, por exemplo, os mencionados por Luckesi (1991);

na verdade, objetivam mostrar alguns projetos que tiveram, ou ainda estão tendo, certo impacto nas escolas, especialmente nas de 1º grau, sem, entretanto, a pretensão de exaurir o assunto, que, sem dúvida, merece ser submetido a uma meta-avaliação.

A evolução da avaliação educacional no Brasil com o objetivo de verificar a eficiência de professores, currículos, programas e sistemas, além de possibilitar a identificação de diferentes tendências, sobretudo quanto ao desempenho educacional, entre outros aspectos, como seria desejável (VIANNA, 1992), ainda está para ser pesquisada e analisada. A partir da década de 60, e ao longo dos anos seguintes, pode-se constatar que alguma coisa importante começou a ser realizada, ainda que de forma algo incipiente, mas revelando um esforço para proceder de acordo com orientação metodológica, especialmente com base em fontes norte-americanas.

A avaliação no contexto educacional brasileiro é quase sempre promovida por órgãos governamentais a nível federal – Ministério da Educação – ou a nível estadual, através das Secretarias de Estado, que, por falta de estrutura, muitas vezes solicitam a colaboração de outras instituições, universidades ou fundações públicas e privadas. Ao contrário do que ocorre nos Estados Unidos, em que as universidades assumem a iniciativa de projetos de avaliação, ainda que com financiamento externo, ou o que se passa na Inglaterra, em que fundações privadas, contratadas para prestação de serviços, realizam pesquisas e avaliação, a nossa situação como será apresentado, é bastante diversa.

A **FUNBEC – Fundação Brasileira para o Ensino de Ciências**, ao iniciar um programa de novos currículos em Física, Matemática, Química, Biologia e Geociências, nos anos 60 e 70, começou, igualmente, uma avaliação de seus programas, contando, para esse fim, com a *expertise* de Hulda Grobman, na área da Biologia. Essa atividade pioneira, entretanto, não teve continuidade em outras instituições, perdendo-se essa rara oportunidade para o desenvolvimento de *know-how* e a formação de capacitações na área de avaliação.

A **Fundação Getúlio Vargas**, igualmente em meados da década de 60, iniciou importante programa de avaliação somativa no Rio de Janeiro, desenvolvendo um instrumento para avaliar a capacitação de crianças ao término do 1º grau na rede oficial, inspirando-se no teste ***Iowa Basic Skills***. O projeto obteve a cola-

boração de Anne Anastasi, Frederick Davis e Robert L. Ebel, que contribuíram para a formação de especialistas brasileiros, ministrando cursos de treinamento em 1965. Razões diversas, inclusive uma radical e abrupta alteração curricular, no antigo Estado da Guanabara (Rio de Janeiro), sem uma prévia avaliação da nova proposta de currículo, provocou a descontinuidade do programa e a consequente dispersão dos grupos de trabalho, que passaram a atuar em outras atividades, muitas vezes estranhas à avaliação.

Ao longo dos anos 70 e na década de 80, face ao processo de massificação das instituições de ensino, especialmente no 3º grau, houve certa intensificação de estudos ligados ao acesso ao ensino superior, sobretudo os relacionados a aspectos psicométricos dos instrumentos de medida e à análise de dados socioeconômicos, como pode ser observado em algumas publicações, especialmente em **Educação e Seleção** (1980-89) e a partir de 1990, em **Estudos em Avaliação Educacional**, ainda que esta revista tenha modificado a orientação editorial da anterior, passando a privilegiar problemas ligados à avaliação em geral, sem se deter, prioritariamente, na problemática da seleção para a Universidade. A década de 70 apresentou, igualmente, interesse, ainda que teórico, na área da avaliação de programas, com a tentativa de disseminação do modelo **CIPP** – contexto, input, processo e produto – desenvolvido por Daniel Stufflebeam e Egon Guba. Alguns poucos trabalhos foram realizados nessa linha metodológica, mas, também, não tiveram prosseguimento, sendo um momento transitório, como habitualmente ocorre na área educacional brasileira.

O **Programa de Expansão e Melhoria do Ensino no Meio Rural do Nordeste Brasileiro – EDURURAL** –, planejado em 1977, na parte referente a estudos e avaliação, esteve a cargo da **Fundação Cearense de Pesquisa**, que se ocupou de aspectos institucionais, e da **Fundação Carlos Chagas**, que centrou suas atividades na avaliação do rendimento escolar (GATTI; VIANNA; DAVIS, 1991). O projeto, com financiamento do Banco Mundial, coletou dados nos anos de 1981, 1983 e 1985, nos Estados do Ceará, Piauí e Pernambuco, por intermédio de provas de Português e Matemática, aplicadas a crianças de 2ª e 4ª séries do Ensino Fundamental, em 603 escolas rurais. Além da qualidade do ensino da escola rural, o projeto **EDURURAL** realizou seis estudos de caso sobre a atuação dessas escolas nos limites das

relações socioeconomicoculturais locais.

O EDURURAL não se limitou a coletar dados sobre o rendimento escolar em Português e Matemática nas 2ªs e 4ªs séries de escolas rurais, que, aliás, mostraram que as crianças daqueles Estados do Nordeste apresentavam uma aprendizagem dos conceitos básicos visivelmente prejudicada, nessas escolas de ensino multisseriado bastante precário (GATTI, 1993). Incluiu, também, no seu projeto, a avaliação de diversas variáveis, como as condições das escolas, perfil dos professores, impacto do treinamento e condições da família, entre outras.

Os seis estudos etnográficos realizados nos três Estados possibilitaram identificar algumas razões explicativas para o baixo rendimento das crianças nas duas áreas curriculares: – rotatividade dos professores, influência política na designação de professores, inconstância na distribuição de livros, material e merenda, que é feita, igualmente, tendo em vista considerações políticas, baixos salários, condições precárias das escolas multisseriadas, infraestrutura curricular deficiente, pouco tempo dedicado ao ensino durante o dia (2.00 a 2.30 horas/dia), frequência irregular dos alunos, doenças das crianças, condições familiares, dificuldade de acesso à escola, ensino baseado na memorização sem significado e passividade induzida do aluno, entre outros fatores, conforme foi caracterizado com realismo esse estado de coisas por Gatti (1994). Um quadro bastante dramático do Nordeste rural, refletido pela situação igualmente trágica de seu ensino, que a avaliação revelou.

Quase no final da década de 80, foi iniciado um amplo programa de avaliação do rendimento de alunos de escolas de 1º grau da Rede Pública em todo o País, por iniciativa do **Instituto Nacional de Estudos e Pesquisas Educacionais – INEP**. Inicialmente, em 1987, essa avaliação (VIANNA; GATTI, 1988; VIANNA, 1989a; VIANNA, 1989b; GATTI; VIANNA; DAVIS, 1991; GATTI, 1993) objetivou identificar, na diversidade do quadro educacional brasileiro, pontos curriculares críticos; verificar o desempenho em aspectos cognitivos básicos de alunos de 1ª, 2ª, 3ª, 5ª e 7ª séries; e subsidiar os professores para uma recuperação de seus alunos em aspectos básicos do currículo escolar. Essa avaliação, ao final, abrangeu uma amostra de 27.455 alunos de 238 escolas em 69 cidades de vários Estados da Federação existentes à época, inclusive no então Território do Amapá. O projeto teve seus desdobra-

mentos, com novas avaliações, a pedido da Secretaria de Estado da Educação do Paraná, envolvendo, inicialmente, alunos de 2ª e 4ª séries de escolas de sete cidades (VIANNA; GATTI, 1988) e depois, uma amostra de alunos de escolas oficiais de 22 outras cidades (VIANNA, 1991), com os mesmos objetivos dos trabalhos propostos pelo INEP.

As avaliações a nível estadual promovidas pelo **Instituto Nacional de Estudos e Pesquisas Educacionais – INEP** – visavam a fornecer às Secretarias de Estado da Educação um conjunto de informações sobre as deficiências da aprendizagem escolar; por outro lado, os projetos procuravam criar, também, condições para que as próprias Secretarias tivessem uma efetiva participação nos assuntos pertinentes à avaliação do rendimento, assim como se envolvessem, ainda, em projetos outros relacionados a programas, sistemas e materiais didáticos (GATTI, 1994). As conclusões do projeto, que mostraram aspectos críticos do ensino de 1º grau em quase 70 cidades das diversas regiões geográficas do país, não podem ser, entretanto, generalizadas, tendo em vista a natureza da amostra, um segmento estatisticamente pouco representativo do universo escolar a nível de 1º grau, e a falta de equalização (*equating*) dos instrumentos. O interessante, e digno de ser destacado, é que esse projeto, iniciado em 1987, mais tarde, em 1991, realizaria pela primeira vez uma **avaliação de escolas do sistema privado de ensino**, mostrando a relação entre condição social e rendimento escolar. O estudo, entretanto, ao comparar o desempenho da escoa privada com o dos alunos das escolas públicas, revelou que nem sempre a escola privada é um mar de excelência e a escola pública também nem sempre é tão ruim quanto se julga, aprioristicamente.

Ainda na década de 80, um novo projeto de avaliação a nível nacional foi realizado por intermédio da Secretaria de Ensino do 2º Grau, do Ministério da Educação, com apoio do Banco Mundial e a colaboração científica da **Fundação Carlos Chagas**, sobre o desempenho escolar de alunos da 3ª série do Ensino Médio (VIANNA, 1991). A partir de uma amostra de 3.972 sujeitos de escolas técnicas federais, escolas estaduais, escolas particulares, escolas oficiais e particulares com habilitação magistério e escolas de formação profissional industrial (SENAI), nas cidades de Fortaleza, Salvador, São Paulo e Curitiba, o projeto identificou variáveis sobre escolaridade e a influência de fatores socioeconô-

micos. Esse estudo, que no fundo é de avaliação do rendimento escolar, apresentou resultados surpreendentes, que eram conhecidos mas não constatados, sobre o alto desempenho das escolas técnicas federais em relação às demais escolas, sem, entretanto, pesquisar as razões desse desempenho.

O Ministério da Educação, utilizando a competência técnica do INEP, no início de 1990, iniciou a implantação de um **Sistema Nacional de Avaliação da Educação Básica – SAEB**, com o objetivo de qualificar os resultados obtidos pelo sistema educacional de ensino público, criar e consolidar competências para a avaliação do sistema educacional (PESTANA, 1992), realizando um trabalho cooperativo entre o MEC e as Secretarias de Estado da Educação. A proposta do SAEB adotou um modelo de estudo de fluxo e de produtividade da UNESCO, com vistas a estudar questões relacionadas com a gestão escolar, competência docente, custo-aluno direto e indireto e rendimento escolar, com base em uma metodologia de amostras relacionadas (WAISELFISZ, 1991), estando em vias de sofrer radical transformação em sua metodologia e nos procedimentos de amostragem, na avaliação do segundo semestre de 1995.

A **Avaliação da Jornada Única em São Paulo**, realizada no segundo semestre de 1992, visou a verificar os efeitos da implantação da Jornada Única no Ciclo Básico, a partir de 1988, em escolas da Grande São Paulo. Ou seja, a avaliação procurou saber, conforme o destaque de Gatti (1992), se o Ciclo Básico/ Jornada Única causou algum impacto, algum diferencial no desempenho e no desenvolvimento intelectual das crianças; se houve diferenciais no desenvolvimento de habilidades básicas que devem ser adquiridas, e se ocorreram alterações na relação aprovação/reprovação/ e evasão.

A partir de 1992, a Secretaria de Educação do Estado de Minas Gerais iniciou a **Avaliação do Sistema Estadual de Ensino de Minas Gerais**[5], envolvendo nesse ano a população de estudantes do Ciclo Básico de Alfabetização e os da 8ª série. Os trabalhos prosseguiram em 1993 com a avaliação da 5ª série do Ensino Fundamental, da 2ª série do Ensino Médio e das 3ª/4ª séries da Habilitação Magistério. Ao todo, cerca de 930.000 estudantes foram avaliados externamente nesses dois anos, sendo levantados dados sobre o desempenho escolar em áreas curriculares, informações socioeconômicas dos alunos, atitudes em relação à ciência (2ª série do

**5** O artigo de Maria Alba de Souza, na revista *Estudos em Avaliação Educacional*, São Paulo, n. 12, 1995, oferece uma visão bastante detalhada do programa de avaliação de Minas Gerais no período de 1992-95.

Ensino Médio), dados relacionados à escola, incluindo elementos sobre os vários currículos, e informações relativas à formação das futuras professoras da escola de Ensino Fundamental (alunas das 3ª/4ª séries da Habilitação Magistério). O programa de avaliação do sistema de ensino de Minas Gerais, integrante de um programa de qualidade da escola, com apoio financeiro do Banco Mundial, teve prosseguimento em 1994, com a repetição do processo nas escolas avaliadas em 1992 e a inclusão de escolas da rede municipal de ensino, abrangendo, no caso, o sistema em 414 municípios. O processo de avaliação em Minas Gerais já começou a produzir seus primeiros resultados, sobretudo em modificações curriculares e na disseminação de centros de atualização de professores em 53 diferentes pontos do Estado (VIANNA, 1992), além da autonomia administrativa e financeira, estando em vias a implantação da autonomia pedagógica, em um processo de qualidade da escola que tem como um de seus pontos centrais a avaliação.

Um programa de avaliação foi realizado em 1991 sobre o **Desempenho da Rede Pública Escolar do Estado de Pernambuco na Área da Linguagem**, em crianças de 1ª a 4ª série, o que possibilitou à Secretaria de Educação do Estado, segundo Buarque *et al.* (1992), a revisão dos conteúdos programáticos das Propostas Curriculares, redirecionamento dos conteúdos nos programas de capacitação de docentes, acompanhamento das classes de alfabetização e o desenvolvimento de um programa de pesquisas.

A nível internacional, no período de 1990/92, houve uma avaliação do desempenho de crianças de 13 anos em Matemática e Ciências, nas cidades de São Paulo e Fortaleza, como parte integrante do projeto da ***International Assessment of Educational Progress – IAEP,*** com a participação de 20 países, sob a coordenação do ***Educational Testing Service*** (Princeton, New Jersey) e o financiamento da ***National Science Foundation – NSF***. A parte relativa ao Brasil ficou sob a responsabilidade da **Fundação Carlos Chagas**, visando o projeto a analisar o desempenho dos estudantes nos vários países participantes e a identificar tipos de ambientes culturais e práticas educacionais associadas a um alto desempenho (VIANNA, 1992). Além das provas de Matemática e Ciências foram desenvolvidos três outros instrumentos – questionário do aluno, da escola e dos pais –, a fim de explorar variáveis associadas ao desempenho educacional

e proporcionar um contexto que possibilitasse compreender os resultados da avaliação. O desempenho das crianças brasileiras, lamentavelmente, foi bastante comprometido, refletindo, assim, a conhecida crise do nosso sistema de ensino da escola de 1º grau. Esse tipo de avaliação costuma, no entanto, com justa razão, ser criticado por vários educadores, sobretudo na Inglaterra, tendo em vista a diversidade curricular, os objetivos do ensino, as metodologias empregadas, entre outros fatores, e inclusive o próprio contexto cultural, em termos sociológicos.

A análise dos vários projetos de avaliação, segundo demonstrou Gatti (1994), evidencia que nem sempre é possível a adoção de um delineamento experimental, que possibilite comparar diferenças de tratamentos; as amostras, muitas vezes, são bastante comprometidas, tendo em vista a carência de estatísticas básicas confiáveis, os procedimentos de aplicação dos instrumentos e coleta de dados nem sempre são merecedores de confiança, apesar das orientações e dos treinamentos a que são submetidos os aplicadores; e a adoção de modelos de avaliação que nem sempre se ajustam ao nosso contexto socioeducacional, havendo necessidade, portanto, de uma adaptação à realidade brasileira. Esses e outros problemas – como, por exemplo, a falta de elementos experientes em avaliação educacional – fazem com que, em muitos casos, o avaliador responsável tome decisões que, em princípio, fogem à ortodoxia doutrinária, mas que resultam da sua experiência e do seu bom senso, face ao imprevisto das situações que surgem, sobretudo em países que não possuem uma tradição de avaliação no seu sistema educacional, como é o caso do Brasil, e em que a realidade é bem diferente da teoria.

A análise da produção científica sobre avaliação educacional publicada em *Cadernos de Pesquisa*, no período de 20 anos (VIANNA, 1992) mostra que o tema é de interesse dos educadores, sendo essa produção bastante diversificada e revelando uma grande preocupação metodológica na abordagem dos vários assuntos, sobretudo na área de rendimento escolar, educação de adultos, treinamento e formação de professores, temas dominantes no contexto da educação brasileira, pelo menos no momento atual. Observa-se, também, nesses artigos e ensaios, um destaque em relação ao emprego de metodologias qualitativas, especialmente estudo de casos, ainda que sejam bastante reduzidos os trabalhos teóricos sobre essas metodologias por autores

nacionais, que se apoiam quase sempre na fundamentação de teóricos estrangeiros, sobretudo norte-americanos e ingleses. A partir dessa análise (VIANNA, 1992), percebe-se, entretanto, que já há uma consciência da importância da avaliação e que a sua prática, apesar de inspirada em outros contextos, e partindo de modelos adaptados à nossa realidade, mas nem sempre de forma cientificamente rigorosa, está começando a criar uma **cultura da avaliação** na sociedade brasileira.

## REFERÊNCIAS BIBLIOGRÁFICAS


BUARQUE, Lair L. *et al*. Avaliação do desempenho da rede pública escolar do estado de Pernambuco na área de linguagem. *Estudos em Avaliação Educacional*, São Paulo, n. 5, p. 95-106, 1992.

CAMPBELL, D. T.; STANLEY, J. C. Experimental and quasi-experimental designs for research on teaching. In: GAGE, N. L. (Ed.). *Handbook of research on teaching*. Chicago, IL: Rand McNally, 1963. p. 171–246.

CRONBACH, L. J. Course improvement through evaluation. *Teachers College Record*, n. 64, p. 672-683, 1963.

GATTI, Bernardete A. O rendimento escolar em distintos setores da sociedade. *Estudos em Avaliação Educacional*, São Paulo, n. 7, p. 95-112, 1993.

_____. Avaliação da jornada única em São Paulo. *Estudos em Avaliação Educacional*, São Paulo, n. 5, p. 85-89, 1992.

_____. Avaliação educacional no Brasil: - experiências, problemas, recomendações. *Estudos em Avaliação Educacional*, São Paulo, n. 10, p. 67-80, 1994.

GATTI, Bernardete A.; VIANNA, Heraldo M.; DAVIS, Claudia. Problemas e impasses da avaliação de projetos e sistemas educacionais: dois casos brasileiros. *Estudos em Avaliação Educacional*, São Paulo, n. 4, p. 07-26, 1991.

LINDQUIST, E. F. *Design and analysis of experiments in Psychology and Education*. Boston, Houghton Mifflin, 1953.

LUCKESI, Cipriano C. *Avaliação dos resultados da aprendizagem escolar:* a prática atual e sua compreensão nas concepções pedagógicas. 1991. Tese (Doutorado) – Pontifícia Universidade Católica de São Paulo, São Paulo, 1991.

MADAUS, G. F.; AIRASIAN, P. W.; KELLAGHAN, T. *School effectiveness:* a reassessment of the evidence. New York: McGraw-Hill Book, 1980.

MADAUS, G. F.; STUFFLEBEAM, D. L.; SCRIVEN, M. S. Program evaluation: a historical overwiew. In: MADAUS, G. F. *et al*. (Ed.). *Evaluation Models:* viewpoints on educational and human services evaluation. Boston: Kluwer-Nijhoff, 1993.

NORRIS, N. *Understanding educational evaluation*. London: Kogan Page, 1993.

PESTANA, Maria Inês G. de S. O Sistema Nacional de Avaliação da Educação Básica. *Estudos em Avaliação Educacional*, São Paulo, n. 5, p. 81-84, 1992.

SCRIVEN, Michael. *The methodology of evaluation*: perspectives of curriculum evaluation. Chicago: Rand McNally, 1967. (AERA. Monograph, n.1).

STAKE, Robert E. The countenance of educational evaluation. *Teachers College Record*, v. 68, n. 7, p. 523-540. 1967.

STUFFLEBEAM, Daniel L. The relevance of the CIPP Evaluation Model for Educational Accountability. *Journal of Research and Development in Education*, v. 5, n. 1, p. 19-25. 1971.

VIANNA, Heraldo M. Avaliação do rendimento de alunos de escolas de 1º grau da rede pública: um estudo em 20 cidades. *Educação e Seleção*, São Paulo, n. 19, p. 33-98, 1989a.

_____. Avaliação do rendimento de alunos de escolas de 1º grau da rede pública: um estudo em 39 cidades. *Educação e Seleção*, São Paulo, n. 20, p. 5-56, 1989b.

_____. *Rendimento de alunos das 2ª e 4ª séries das escolas oficiais do estado do Paraná*: um estudo avaliativo. São Paulo: Fundação Carlos Chagas, 1991. Relatório de pesquisa.

_____ Avaliação do rendimento escolar de alunos da 3ª série do 2º grau: subsídios para uma discussão. *Estudos em Avaliação Educacional*, São Paulo, n. 3, 1991.

_____. Avaliação do desempenho em matemática e ciências: uma experiência em São Paulo e em Fortaleza. *Estudos em Avaliação Educacional*, São Paulo, n. 5, p. 9-14 , 1992a.

_____. Avaliando a avaliação: da prática à pesquisa. *Estudos em Avaliação Educacional*, São Paulo, n. 5, p. 55-64, 1992b.

_____. Avaliação do ciclo básico de alfabetização em Minas Gerais. *Estudos em Avaliação Educacional*, São Paulo, n. 5, p. 91-94, 1992c.

_____. Avaliação Educacional nos Cadernos de Pesquisa. *Cadernos de Pesquisa*, São Paulo, n. 80, p. 100-105, 1992d.

VIANNA, Heraldo M.; ANTUNES, Ana Lúcia; SOUZA, Maria Alba de. Desenvolvimento de um programa de avaliação do sistema estadual de ensino: o exemplo de Minas Gerais. *Estudos em Avaliação Educacional*, São Paulo, n. 8, p. 5-37, 1993.

VIANNA, H. M.; FRANCO, Gláucia T. Avaliação do rendimento de alunos de escolas do 1º grau da rede privada: pontos críticos e convergência. *Estudos em Avaliação Educacional*, São Paulo, n. 7, p. 113-132, 1993.

VIANNA, Heraldo M.; GATTI, Bernardete A. Avaliação do rendimento de alunos de escolas de 1º grau da rede pública: uma aplicação experimental em 10 cidades. *Educação e Seleção*, São Paulo, n. 17, p. 5-52, jan./jun. 1988a.

_____. Avaliação do rendimento de alunos de 2ª e 4ª séries de escolas oficiais do estado do Paraná. *Educação e Seleção*, São Paulo, n. 18, p. 5-62, 1988b.

WAISELFISZ, Júlio J. O Sistema Nacional de Avaliação Educacional de 1º Grau. *Estudos em Avaliação Educacional*, São Paulo, n. 4, p. 65-72, 1991.

WORTHEN, B. R.; SANDERS, J. R. *Educational Evaluation: alternative approaches and practical guidelines*. New York, London: Longman, 1987.

# MEDIDA DA QUALIDADE EM EDUCAÇÃO: APRESENTAÇÃO DE UM MODELO[1]

O problema da qualidade em educação é uma preocupação da sociedade como um todo. Subjacentemente, o conceito de responsabilidade educacional – *educational accountability* – permeia essa inquietação, ou seja, havendo pessoal qualificado, condições materiais, instrumental instrucional, metodologias e estratégias adequadas, a educação formal deveria ser, necessariamente, de boa qualidade. Isso, entretanto, nem sempre ocorre, pois uma complexa rede de variáveis atua no processo e cria um quadro de elementos interferentes que determinam níveis diversos de excelência educacional.

Algumas indagações surgem de imediato: como conceituar qualidade em educação? Será possível uma definição operacional de qualidade em educação a fim de mensurá-la com adequação? O problema precisa ser analisado e discutido com a participação da comunidade educacional e de elementos da sociedade. A medida da qualidade em educação, entretanto, não pode ficar restrita apenas ao desempenho escolar, necessita, também, verificar outras variáveis que se associam e condicionam o rendimento escolar. O que as crianças fazem na escola, o que os professores procuram transmitir aos seus alunos e o que os livros didáticos

**1** Artigo publicado na revista *Estudos em Avaliação Educacional*, São Paulo, n. 2, p. 99-104, jul./dez. 1990.

apresentam refletem expectativas culturais e educacionais na sociedade, bem como seus valores e seus objetivos sociais e econômicos. Assim, é impositivo verificar em que medida a interação dessas variáveis contribui para a qualidade da educação.

É necessário reiterar que a avaliação da qualidade da educação não se limita apenas à verificação do rendimento escolar, que é um momento na caracterização dessa qualidade. O desempenho dos estudantes em pesquisas da qualidade da educação é melhor compreendido e interpretado quando se levantam informações sobre o tipo de ensino que recebem, os procedimentos que vivenciam em sala de aula e no colégio, e ainda sobre as características ambientais da família que determinam o seu comportamento. Assim, a pesquisa sobre a qualidade da educação precisa caracterizar o *contexto nacional* em que o processo educacional se desenvolve, identificar criticamente *os fatores não diretamente ligados à escola* que afetam a educação e analisar a ação da *escola em termos de entrada, processo e produto*.

A avaliação da qualidade da educação deve, necessariamente, partir de uma análise do *contexto nacional*, envolvendo as características da população, os seus valores culturais, os investimentos financeiros em educação e a organização das escolas. As *características da população* não se devem limitar a estatísticas demográficas, mas apresentar e discutir os vários níveis de educação, o processo de transformação da economia e a composição da força de trabalho, assim como suas tendências. Os *valores culturais* da sociedade precisam ser identificados, destacando-se, particularmente, a problemática da maior ou menor valorização da educação, o papel da educação no desenvolvimento individual e na formação profissional, as oportunidades educacionais oferecidas pela sociedade e o grau de universalização da educação; além disso, problemas cruciais como o do *status* do professor na sociedade e a responsabilidade da família na educação necessitam também ser considerados. Os *investimentos financeiros em educação* não podem deixar de ser considerados como uma das importantes variáveis que influenciam na qualidade da educação, juntamente com a alocação de recursos humanos qualificados para a área educacional. Finalmente, no quadro do contexto nacional, a avaliação da qualidade da educação deve considerar a organização do sistema escolar, identificando os diversos tipos de escola, os graus de centralização administrativa, a influência

do processo de seletividade social na escola e a organização do sistema de avaliação, que, dependendo da sua filosofia e da sua estruturação, pode criar diferentes tipos de impedimentos, com amplas repercussões sociais.

A criança, e mesmo o adolescente, na maioria dos casos, passa um tempo extremamente reduzido na escola; desse modo, a fim de avaliar a qualidade da sua educação, é fundamental que sejam considerados os *fatores não diretamente ligados à escola*, que compreenderão, entre outros aspectos, o *status* socioeconômico da família, o nível de educação dos pais, os recursos educacionais no lar, o interesse e a participação dos pais no processo educacional, as atividades educacionais fora da escola, as atividades de lazer e sociais (televisão, esportes e interação de grupos) e, ainda, uma análise das atitudes e das aspirações dos estudantes.

A investigação sobre a qualidade da educação envolve, naturalmente, uma discussão aprofundada da variável '*escola*', que, no presente modelo, deve sofrer uma abordagem em termos de entrada, processo e produto. A partir dessa perspectiva, para a configuração do fator '*entrada*', é desejável a identificação e análise das variáveis tamanho e tipo de escola; extensão do ano letivo e da jornada escolar; tamanho, características e experiência do pessoal docente; qualidade das instalações escolares; organização dos programas escolares; e, finalmente, a participação dos pais na vida escolar.

A parte do modelo mais diretamente relacionada à sala de aula, ao que é ensinado e como é ensinado, no seu conjunto, engloba a variável '*processo*', que se centra na avaliação do *currículo* e das *práticas instrucionais*. A avaliação do currículo visa, particularmente, a análise dos objetivos e dos conteúdos programáticos, procurando identificar em que medida são destacados fatos, conceitos e/ou habilidades complexas. Ainda na variável '*processo*', o conjunto das práticas instrucionais procura identificar o tipo de instrução ministrada (em grandes e pequenos grupos e/ou instrução programada); a ocorrência de aulas expositivas, demonstrações e/ou discussões; o uso de material didático (impressos e equipamentos); a natureza das tarefas a serem realizadas em casa; e, finalmente, a crítica das práticas docimológicas, ou seja, a análise do planejamento e da execução da avaliação escolar.

O presente modelo completa-se com a avaliação da variável '*produto*', concentrada no desempenho escolar e na formação de atitudes associadas ao processo educacional, momento em que serão empregados diferentes instrumentos de mensuração escolar, inclusive, testes de escolaridade e questionários de atitudes.

Um projeto de avaliação da qualidade da educação, dessa forma, deve relacionar o desempenho dos estudantes a contextos culturais e a práticas educacionais associadas ao rendimento escolar. Surge, assim, um grave problema a ser discutido – a validade dos instrumentos, particularmente a validade dos testes de escolaridade.

O uso de testes sem a observância da teoria das medidas contribui para que sejam negligenciadas importantes características dos instrumentos aplicados à mensuração dos atributos humanos. Diferentes testes possuem diferentes virtudes e não existe um teste que seja o melhor para todos os propósitos. Um teste é sempre construído com uma determinada finalidade, não havendo, portanto, um único instrumento que seja capaz de medir de modo globalizado as diferentes dimensões do homem. Assim, ao construir um teste, é necessário que se identifique o objetivo do instrumento e em que situação concreta será utilizado. Um teste para uso em sala de aula, com a finalidade de promover uma *avaliação formativa*, não tem, certamente, as mesmas características de um outro empregado para fins de seleção, quando se processa uma *avaliação somativa*. Um teste serve, ainda, para um determinado contexto e para um fim especifico, não tendo a mesma utilidade em outros contextos e para outros fins.

As questões relativas à validade dos testes educacionais são complexas e, frequentemente, sujeitas a controvérsias, que, entretanto, precisam ser enfrentadas a fim de solucionar os problemas relativos à construção dos instrumentos de medida, à sua análise empírica e ao julgamento do seu valor como elementos capazes de medirem, eficientemente, os atributos definidos no planejamento.

O processo de avaliação de um teste possui uma lógica própria, ressaltando-se, entretanto, que a validação de um teste depende do uso que dele se pretenda fazer. Um teste nunca possui um único e exclusivo objetivo; desse modo, não se pode assegurar que um teste seja válido em termos gerais. A validação de um instrumento de medida educacional está, pois, relacionada com a interpretação que ao mesmo se pretenda dar.

A validade de um instrumento de medida está associada à concretização dos seus objetivos. Assim, no caso específico de um teste de desempenho escolar, a validade é positivada na medida em que esse instrumento realiza aquilo a que se propuseram seus construtores. O conceito de validade, entretanto, é multifacetado, não sendo possível dizer, *a priori*, se um teste é ou não válido: um teste pode ser válido para um currículo e não o ser para outro; um teste pode ser válido para um grupo, mas não para outro. A validade não é, pois, um atributo que se possa apresentar em termos gerais; é sempre específica a um instrumento, a um curso, a um currículo, a um professor e a um grupo de indivíduos com características bem definidas.

O processo de validação presta-se a diferentes interpretações muitas vezes conflitantes entre si, sendo necessário, por conseguinte, maiores esclarecimentos sobre sua natureza. Valida-se não propriamente o teste mas a interpretação dos dados decorrentes de um procedimento específico. Um mesmo teste pode ser usado para diferentes fins e a cada aplicação do instrumento pode corresponder, portanto, uma interpretação dos resultados. Ora, considerando que cada interpretação tem seu próprio grau de validade, não é possível afirmar que um teste é válido em geral, pois a validade é sempre específica. Um teste, em síntese, possui muitas validades; consequentemente, há necessidade de aprofundar e estender os estudos de validação para que se possa determinar as diversas dimensões da sua validade.

O problema da validação dos instrumentos deve receber tratamento prioritário no processo investigativo da qualidade da educação, conforme a análise anterior. O uso indiscriminado de instrumentos de avaliação pode levar a conclusões inteiramente falaciosas, com amplas repercussões. Nesse momento, cabem algumas indagações referentes ao contexto educacional brasileiro: – os instrumentos usados para fins de avaliação educacional possuem validade empiricamente verificada? Os instrumentos usados pelos professores possuem validade curricular? Os instrumentos empregados no dia a dia da escola possuem validade de conteúdo? E, finalmente, uma pergunta dolorosa: – os altos índices de reprovação nas primeiras séries do 1º grau não decorreriam em parte da falta de validade dos instrumentos usados na avaliação? São questões que precisariam ser respondidas no processo de investigação da qualidade da educação.

## ANEXO

### MEDIDA DA QUALIDADE EM EDUCAÇÃO
### ESQUEMA DE UM MODELO

**1.0.** CONTEXTO NACIONAL
**1.1.** Características da população
- **1.1.1.** Estatísticas demográficas
- **1.1.2.** Níveis de educação
- **1.1.3.** Transformações da economia
- **1.1.4.** Força de trabalho

**1.2.** Valores culturais
- **1.2.1.** Valorização da educação
- **1.2.2.** Desenvolvimento individual
- **1.2.3.** Formação profissional
- **1.2.4.** Oportunidades educacionais
- **1.2.5.** Universalização da educação
- **1.2.6.** *Status* do professor
- **1.2.7.** Responsabilidade da família na educação

**1.3.** Investimentos financeiros
- **1.3.1.** Alocação de recursos humanos

**1.4.** Organização da escola
- **1.4.1.** Tipos de escola
- **1.4.2.** Centralização administrativa
- **1.4.3.** Seletividade social
- **1.4.4.** Sistema de avaliação

**2.0.** FATORES NÃO DIRETAMENTE LIGADOS À ESCOLA
**2.1.** *Status* socioeconômico da família
**2.2.** Nível de educação dos pais
**2.3.** Recursos educacionais no lar
**2.4.** Participação dos pais no processo educacional
**2.5.** Atividades educacionais fora da escola
**2.6.** Atividade de lazer e sociais
**2.7.** Atitudes e aspirações dos estudantes

**3.0.** ESCOLA
**3.1.** Entrada
- **3.1.1.** Tamanho e tipo da escola
- **3.1.2.** Extensão do ano letivo e da jornada escolar
- **3.1.3.** Tamanho, características e experiência do corpo docente

**3.1.4.** Qualidade das instalações escolares
**3.1.5.** Organização dos programas escolares
**3.1.6.** Participação dos pais na vida escolar
**3.2.** Processo
**3.2.1.** Currículo
**3.2.1.1.** Análise de objetivos e conteúdos programáticos
**3.2.2.** Práticas instrucionais
**3.2.2.1.** Tipos de instrução
**3.2.2.2.** Aulas expositivas, demonstrações e/ou discussões
**3.2.2.3.** Uso de material didático
**3.2.2.4.** Tarefas de casa
**3.2.2.5.** Avaliação escolar
**3.3.** Produto
**3.3.1.** Desempenho escolar
**3.3.2.** Formação de atitudes

# AVALIAÇÃO DE PROGRAMAS EDUCACIONAIS: DUAS QUESTÕES[1]

*Program and project evaluation is a human activity that is as old as mankind. It is used to make choices and decisions. It motivates change. It is used to develop understanding. It is used to justify investments of resources. In many ways program and project evaluation has been and will continue to be embedded in human cognition and behavior.*

*It is fair to say that every program or project evaluation is unique in many ways and consequently that the practice of evaluation is complex.*
JAMES R. SANDERS[2]

## 1. INTRODUÇÃO

A avaliação educacional, no contexto brasileiro, começou a desenvolver-se tardiamente, em meados dos anos 60, ainda que com a quase total centralização nos processos de medida, situação esta que subsiste nos tempos atuais. Apesar desse relativo desenvolvimento, foi ignorado, por alguns segmentos, o fundamento teórico da avaliação; a subjacente teoria da avaliação em

**1** Artigo publicado na revista *Estudos em Avaliação Educacional*, v. 16, n. 32, p. 43-55, jul./dez. 2005.

**2** In: Kellaghan e Stufflebeam (2003).

que se baseia a ação avaliativa, aspecto este que será ressaltado no presente trabalho. É forçoso reconhecer, entretanto, que já se desenvolveu relativo domínio na avaliação da aprendizagem, na avaliação de competências, na avaliação para fins de seleção e na avaliação de atitudes, começando a ocorrer, no momento fluente, maior domínio da avaliação institucional. Observa-se, quando alguns relatórios são examinados, que a avaliação de programas ainda não está adequadamente definida em todas as suas dimensões, sendo, muitas vezes, confundida com a frequentemente chamada avaliação do desempenho.

A literatura existente sobre avaliação de programas ressalta diferentes aspectos, como a sua relevância social, as questões técnicas relacionadas com os diversos tipos de validade, a importância da disseminação dos resultados e seus efeitos, entre outras considerações que são abordadas por diferentes autores, inclusive na importante obra editada por Kellaghan e Stufflebeam (2003). Os objetivos do presente artigo limitam-se, entretanto, a duas questões que se referem à natureza do processo de avaliação e ao que se espera dos esforços desenvolvidos em uma avaliação de programa.

## 2. AVALIAÇÃO DE PROGRAMAS: UM PROCESSO DEMOCRÁTICO

O problema da avaliação como um processo democrático foi discutido com bastante amplitude por House e Howe (2003) e permite considerações relativas a esse assunto nem sempre explorado por teóricos e praticantes da avaliação. O trabalho de avaliar é uma atividade de equipe e, como tal, pressupõe uma ação sem conflitos internos, visando a garantir a validade das conclusões a que a equipe possa chegar na consecução dos seus objetivos. É bom destacar que toda avaliação, inclusive a de programas, tem inserido em seu contexto interesses de diferentes pessoas que refletem diversidade de valores, várias visões de mundo e, consequentemente, posicionamentos diversos, traduzidos, por vezes, em situações de conflito, que fragilizam a equipe e perturbam o andamento dos trabalhos. Seria bom lembrar, nesse momento, que, segundo Merleau-Ponty (2004), "... Cada ser é um só, e ninguém pode dispensar os outros... Não há vida em grupo que nos livre do peso

de nós mesmos, que nos dispense de ter uma opinião...” Fica configurada, dessa forma, a responsabilidade do avaliador na condução de uma equipe de especialistas, com pluralidade de pensares, que procura avaliar um programa.

Assim, como deixam perfeitamente claro House e Howe (2003), é fundamental que se considere quais os interesses, valores e visões das várias audiências[3] em relação ao programa a ser avaliado. Ressalte-se, entretanto, que somente serão levados em conta os interesses, os valores e as visões que são efetivamente importantes para o programa.

**3** Entende-se como audiências (*stakeholders*, em inglês) todas e quaisquer pessoas envolvidas ou afetadas pela avaliação: estudantes, pais/responsáveis, professores, administradores, orientadores, psicólogos, associações de pais e mestres, futuros empregadores, membros da comunidade e outros que tomem decisões que afetem a educação do estudante.

A avaliação, conforme o destaque anterior, é um trabalho de grupo, que exige a integração dos seus diversos elementos. Podem ocorrer, entretanto, distorções no pensar e no agir de alguns elementos que impedem a integração do grupo, por ausência de um diálogo fluente e aberto que conduza a decisões sensatas e adequadas à situação. Ainda que difícil, deve-se levar a equipe de avaliação a um consenso, que vai possibilitar ao grupo agir de forma coesa, com a superação de conflitos.

É importante ressaltar o papel do avaliador como elemento aglutinante do grupo, numa avaliação realizada segundo a perspectiva democrática, sendo a ele vedada a sua identificação, ainda que não explícita, com subgrupos que possam vir a surgir na equipe. A ocorrência desses subgrupos deve ser evitada de forma enfática, impondo-se, ainda, por outro lado, que os possíveis desequilíbrios no grupo, gerados por manifestações de autointeresse, sejam igualmente controlados, em respeito ao objetivo maior do projeto.

A avaliação de programas, segundo a perspectiva democrática, pressupõe a participação das várias audiências no desenvolvimento da sua estrutura. Esse aspecto exige, necessariamente, que sejam definidas por antecipação as regras e os procedimentos para esse fim. Isso significa a realização de reuniões prévias, organizadas de acordo com as características das audiências, para que elas possam manifestar suas preocupações e suas ideias relativamente à avaliação e ao respectivo projeto, por via de um processo de interação destituído de formalismos. As audiências, portanto, participam das deliberações relativas ao projeto e, no decorrer dessas reuniões interativas, o avaliador responsável pela equipe procura identificar os reais interesses do grupo, levado-o a uma ação reflexiva. Durante essas reuniões, para fins de deliberação, o avaliador que lidera a equipe pode testar a consis-

tência dos critérios adotados no projeto de avaliação e, ao longo de sua realização, comprovar, igualmente, a coerência dos dados parcialmente coletados, para que, ao término dos trabalhos, sejam válidas as conclusões acerca do programa avaliado.

## 3. ELEMENTOS RESULTANTES DE UMA AVALIAÇÃO DE PROGRAMAS

A indagação sobre quais seriam os elementos que resultariam das várias ações em uma avaliação de programa admite múltiplas respostas, se considerarmos o grande número de modelos existentes, como pode ser constatado na obra de Madaus *et al.* (1993); desse modo, para responder à questão inicialmente proposta, o presente artigo se baseará em trabalho de Stake (1973), que teve larga repercussão e, decorridos mais de trinta anos, ainda pode servir de orientação para aqueles que se dedicam à avaliação de programas.

A expectativa, segundo o ponto de vista a ser seguido, é de que o relatório da avaliação apresente os objetivos definidos em razão das informações coletadas nas reuniões interativas com as audiências interessadas. Algumas informações são significativas para certos grupos de pessoas, mas essas mesmas informações não atendem aos interesses de outros segmentos das audiências consultadas. O que importa aos responsáveis pela formulação do programa pode não corresponder aos aspectos que são de interesse imediato dos pais de alunos, por exemplo. Assim, é fundamental ter em mente a especificidade dos vários grupos integrantes das audiências, na fase de elaboração de relatórios parciais, no decorrer do processo de avaliação e no momento de estruturação do relatório final.

O documento de disseminação dos resultados da avaliação deve ter um caráter eminentemente descritivo e levar em consideração o público a que se destina. Ele pode não ser de interesse imediato para os que participaram da implementação do programa no seu dia a dia (professores, técnicos escolares, administradores, entre outros), mas ser de grande valia para pesquisadores educacionais e especialistas em avaliação; desse modo, fica definido que na divulgação dos dados de uma avaliação de programa, para que ela tenha impacto, é necessário que a cada tipo de audiência corresponda um relatório específico, variando de um

relatório técnico para especialistas, com as suas complexidades estatísticas, quando for o caso, a um folheto de divulgação dos elementos mais representativos para a sociedade.

Ao ser caracterizada a avaliação como um processo democrático, foi reiterada a necessidade de reuniões interativas com as diversas audiências, ao longo do seu processo. É claro que essas reuniões se devem caracterizar por um intercâmbio de ideias seminais, que fundamentarão o processo de avaliação. A avaliação não é um produto que nasce feito e definido em todas as suas características, ele se constrói com base na troca de pontos de vista, que geram discussões que passam a integrar todo o processo. É igualmente necessário aproveitar a temática desses diversos falares que enriquecem a avaliação e que necessitam ser registrados. Essas discussões, às vezes, são anteriores ao programa e, com grande frequência, outras ocorrem após a sua implantação. O avaliador, nesse caso, ver-se-á obrigado a fazer suposições a respeito do programa, submetendo as suas considerações a rigorosa análise do grupo. As discussões ocorridas constituem um acervo a preservar e são sempre úteis, inclusive no caso de futuras avaliações.

A metodologia de uma avaliação apresenta elementos caracterizados por Stake (1973) como antecedentes que devem ser documentados nos diversos relatórios apresentados às audiências. É necessário que se detalhe a especificação do programa de avaliação, apresentando às audiências os elementos seguintes: 1) o que é o programa e em que consiste; 2) o que ocorreu durante a sua implementação (em sala, em laboratórios ou em outros locais); 3) o que foi tentado, ainda que não tenha sido bem sucedido; e, finalmente, 4) os dados do resultado do programa. Ao relatório incorporam-se também outros importantes subsídios, como: 1) aquilo que se pretendeu e 2) o que foi efetivamente constatado, acrescentando-se outros elementos que, eventualmente, possam ser úteis para a formulação de juízos de valor pela equipe de avaliação. Entre os elementos caracterizados por Stake (1973) como antecedentes, incluem-se ainda: 1) dados demográficos e escolares dos estudantes; 2) características gerais e profissionais dos professores; 3) conteúdos do currículo; 4) material utilizado durante a instrução; 5) descrição física da instituição; 6) organização da escola; e 7) análise do contexto da comunidade em que o programa se desenvolveu.

Ainda dentro da linha apresentada por Stake (1973), a avaliação de programa se preocupa em medir diferentes dimensões, envolvendo habilidades, compreensão e capacidade de interpretar conceitos, por exemplo, cujos resultados devem ser apresentados e interpretados nos relatórios parciais e no documento final. Essas mensurações, ocorridas em diferentes momentos da avaliação do programa, procuram apresentar as suas consequências, ou seja, os seus resultados. Além desses resultados, muitas vezes, no decorrer do projeto, são verificadas atitudes e habilidades motoras. É necessário que se ressalte, quanto às mensurações, que elas não podem ficar restritas ao esquema bastante simplista do pré e pós-teste, devendo, ao contrário, ser adotado um esquema de avaliação formativa.

O relatório de avaliação apresenta, dessa forma, os efeitos do programa em relação aos alunos, porém não fica restrito a esse segmento. Oferece, também, informações relacionadas ao impacto sofrido pelos professores e pela própria instituição em que o programa está sendo aplicado. Essas informações devem ser levantadas por intermédio de múltiplas coletas, com o uso de diferentes tipos de instrumentos, e não podem ficar limitadas à aplicação de um único teste, questionário ou a um simples julgamento de professores, como ocorre em frequentes avaliações.

Ao apresentar o relatório, é necessário que o avaliador responsável pela equipe de trabalho insira uma descrição da filosofia que serviu de base para a definição e estruturação do programa que está sendo objeto da avaliação ou, resumindo, a fundamentação teórica do programa, que não pode ser desconhecida pelos que avaliam, mas é muitas vezes ignorada por quem implementa o projeto. É preciso levar em conta, por outro lado, que existem diferentes posicionamentos filosóficos em relação à educação, sendo necessário que se caracterize, no documento de disseminação da avaliação, qual a filosofia que serviu de base para a definição e a construção do programa.

O avaliador, ao considerar o conjunto das atividades que serão empreendidas pela sua equipe, necessita considerar o corpo do programa na sua inteireza e identificar os aspectos que terão maior destaque pela importância no todo a ser avaliado. Nem tudo é incorporado ao conjunto das preocupações do avaliador. Há aspectos que, por seu significado, configuram realmente o programa e, assim, merecem ser considerados; são

partes que, por serem essenciais ao programa, vão exercer influência no seu êxito.

Stake (1973), com razão, chama a atenção para o fato de que o avaliador não deve procurar identificar, na análise do programa, objetivos referentes a comportamentos. A sua preocupação centra-se, especialmente, nos objetivos da instituição que adotou o programa, nos dos pais/responsáveis em relação ao programa, assim como nos objetivos dos alunos, dos professores e nos objetivos que por ventura outros membros da instituição interessados no programa possam ter. Esses objetivos são considerados durante o transcurso da avaliação, e as várias audiências ficarão a par desses objetivos surgidos aos poucos e identificados no decorrer do processo de avaliação. É preciso ficar claro, portanto, que a avaliação não parte de objetivos, mas os vai identificando no decorrer do processo.

A necessidade de coletar informações ao longo da avaliação é encarecida por Stake (1973), que, entre outras informações, destaca as que se referem 1) a procedimentos instrucionais, 2) a estratégias de ensino adotadas, e 3) a diferentes meios (e multimeios) empregados pelos professores. Esses elementos podem ser levantados por intermédio de entrevistas e pela utilização das múltiplas técnicas de observação (VIANNA, 2000). Ao serem realizadas as observações é necessário que se caracterizem os elementos que comprovam o efetivo desempenho dos professores e dos alunos em sala de aula. Além do mais, é importante que, por intermédio da observação, seja constatada a interação professor/aluno, a ser objeto de análise pela equipe.

Ao realizar a avaliação, além das características ligadas ao nível de competência dos alunos, precisa ser considerada a sua situação social, variável esta que pode concorrer para explicar o desempenho escolar. A situação social, segundo Stake (1973), pode ser apresentada sob a forma de um sociograma[4], sendo útil esse dado para mostrar os diferentes tipos de contato social, caso se realizem avaliações sucessivas. Outro aspecto a considerar refere-se ao contexto social, ou seja, à relação entre a comunidade e a escola. Ambas devem ser apresentadas com o máximo de detalhes para que, na hipótese de comparações e generalizações, o avaliador possa decidir sobre a efetividade desses procedimentos.

**4** Para uma análise dos métodos sociométricos e sua fundamentação, ver Merleau-Ponty (1990).

É preciso atentar para o fato de que, ao ser adotado o esquema ora apresentado, com base na orientação de Robert L.

Stake, é imprescindível a fixação e definição de padrões para que se possa fazer julgamento de valor sobre a qualidade efetiva do programa. Ao serem definidos esses padrões devem ser levados em conta valores, crenças e múltiplas exigências sociais em relação aos vários tipos de aprendizagem e às diferentes situações da escola. É reconhecidamente sabido por todos que se dedicam à avaliação que esses padrões são difíceis de estabelecer. Assim, tomando-se uma situação hipotética: avaliação de um programa de Matemática ou um outro de Literatura Brasileira para o ensino médio, seria certamente necessária uma consulta significativa para saber a opinião de professores universitários e de professores atuantes nos dois últimos anos do ensino básico, a fim de que seja possível estabelecer padrões de qualidade relevantes para a programação educacional dessas duas áreas curriculares. Ressalte-se, entretanto, que, nessa situação, nem sempre é possível chegar a um resultado consensual, mas o empreendimento precisa ser tentado.

Qual o impacto do programa? Uma avaliação procura determinar em que medida o programa teve algum significado para escola e para o sistema ao possibilitar outras experiências e mudanças de comportamento. Um novo programa, no dizer de Stake, deve oferecer oportunidades para que modificações sociais igualmente ocorram. Por exemplo, qual o impacto da visita dos alunos da 8ª série de uma escola estadual à Bienal do Livro? Qual o impacto das palestras de professores universitários sobre a importância da teoria da relatividade em um novo programa de física para os alunos de uma escola de ensino médio? Qual o impacto provocado nas primeiras séries do ensino fundamental em decorrência da introdução de um programa sobre a arte de contar estórias para professoras? Ou seja, o programa a ser avaliado proporcionou novas oportunidades aos alunos? Houve, realmente, alguma mudança nos alunos, nos professores e na própria escola? Sem dúvida, essas são importantes indagações a serem propostas pelo avaliador ao iniciar o seu trabalho, conforme as colocações de Stake (1973).

Um programa, em princípio, deve proporcionar ganhos aos alunos. Muitas vezes, entretanto, os avaliadores se decepcionam porque os ganhos esperados são reduzidos. Uma primeira explicação para o problema estaria, muito possivelmente, na ausência de instrumentos capazes de constatar diferenças nos diversos

tipos de ganhos no transcurso de um programa. Os instrumentos ora empregados podem ser válidos para a medida de diferentes tipos de atitudes e a identificação de diversos níveis de desempenho, mas não refletem, necessariamente, o impacto dos programas. Essa é uma problemática bastante séria, porquanto, salvo as exceções de sempre, poucas ou raras são as instituições que se dedicam a criar instrumentos sensíveis à medida do impacto provocado pelo desenvolvimento de um programa; consequentemente, como atesta Stake (1973), é difícil apresentar um quadro completo do que está efetivamente ocorrendo no âmbito das instituições educacionais.

A caracterização dos efeitos de um programa é outro aspecto a considerar. Algo acontece, sem sombra de dúvida, nas salas de aula, algo além do tradicional detalhamento dos programas curriculares e das possíveis expectativas de professores, mas esse algo apresenta dificuldades para a tarefa do avaliador, mesmo que seja um profissional bastante experimentado em seu *métier*. Um caminho provável para tentar identificar alguma coisa sobre esse algo desconhecido, que todos sabem existir, segundo a colocação de Stake (1973), estaria em analisar as habituais críticas aos programas educacionais; os comentários, muitas vezes azedos, sobre as metodologias utilizadas; e, ainda, as críticas ao material didático empregado, entre outros aspectos. Uma outra alternativa possível, proposta por Stake (1973), consistiria em fazer uma análise dos dados de diferentes pesquisas educacionais, dados estes que permitirão ao avaliador identificar variáveis úteis para a avaliação e para a solução de seus possíveis problemas.

Todo projeto de avaliação tem um custo que não é apenas financeiro, mas que precisa ser estimado e o seu orçamento controlado, tendo em vista solicitações e novas ideias que possam gerar despesas não previstas. Outros custos existem, como os de pessoal, considerando que uma avaliação de programa exige uma equipe capacitada, cujo custo nem sempre é possível traduzir em termos quantitativos, mas que não deixa de ser um custo, e que a avaliação de programas demanda um tempo razoavelmente longo para a sua concretização. Por outro lado, há o custo relativo ao aluno, que se vê envolvido numa atividade nem sempre interessante para ele. A sua participação significa horas extras de trabalho, aumento de tensão e a consciên-

cia de que está sendo observado por um estranho ao seu grupo, e por mais que se tente explicar que a avaliação é do programa, o aluno sempre acredita ser o objeto imediato da avaliação. Todo esse custo deve ser ponderado, conforme o destaque de Stake (1973), vindo a constituir-se em uma variável cujo valor, no final, nem sempre pode ser estimado com precisão.

Ao ser realizada a avaliação de um programa é imprescindível que se estabeleçam os vínculos que possam existir entre aquilo que foi pretendido e o que foi realmente observado. A análise dessa convergência vai mostrar ao avaliador se pode considerar como aceitável ou não essa ocorrência. Nesse ponto, o avaliador se coloca diante de uma grande interrogação: o que causa realmente o quê? ou, em outras palavras, quais as ações que correspondem a determinados resultados? Fica claro que há necessidade de estudar de que modo as coisas variam simultaneamente, a fim de que se possa estabelecer relações de causa e efeito. A partir dessas relações, comparações podem ser feitas; atente-se, entretanto, para o fato de que a covariância muitas vezes é pouco frequente, ainda que não seja impossível de ocorrer.

Avaliar está associado a julgamentos de valor e sobre isso parece haver razoável consenso, na medida em que é possível chegar a um consenso no campo da educação. Algumas indagações novamente se impõem, como fez Stake (1973): qual o valor do programa? qual o valor das informações coletadas? ou, qual o valor dos objetivos do programa? Ao fazer uma avaliação de programa, o avaliador procura estabelecer, primeiramente, questões de mérito e, a seguir, apresentar os problemas que foram revelados no decorrer do processo de avaliação. Reitere-se, mais uma vez, que na avaliação de programas não se visa a chegar a uma posição consensual, o que é quase impossível, considerando que alunos, professores e administradores podem ter visões inteiramente diferentes sobre um mesmo objeto; assim, a partir de múltiplas fontes, com as mais variadas percepções, o avaliador vê-se na contingência de apresentar juízos de valor que nem sempre coincidem com os que foram manifestados por outras pessoas. A relevância de uma avaliação, consequentemente, vai decorrer do bom senso do avaliador, da sua *expertise* e da sua capacidade de interagir com diferentes audiências interessadas[5].

**5** Para uma visão mais ampla da avaliação de programas em países com larga experiência nessa atividade, ver os artigos de Jean A. King, Alice Dignard e John M. Owen, em Kellaghan e Stufflebeam, 2003, v. II, p.721-768, que tratam da temática, respectivamente, nos Estados Unidos, no Canadá e na Austrália.

## 4. CONSIDERAÇÕES FINAIS

As partes envolvidas na avaliação de um programa precisam partir, necessariamente, de um processo de negociação, conforme destaque anterior, a fim de que os seus interesses sejam definidos e a avaliação possa atingir os objetivos propostos. A negociação vai possibilitar que os grupos envolvidos, atendendo à dinâmica de um processo democrático, definam suas concordâncias e eliminem possíveis situações de conflito, prejudiciais ao trabalho.

O impacto de um programa de avaliação está diretamente ligado ao sucesso da disseminação dos resultados. Uma política de divulgação dos resultados precisa considerar os diferentes grupos interessados para que, em função de suas características, possam ser selecionadas as informações que correspondem à diversidade dos vários interesses. Assim, a validade de uma avaliação depende grandemente da disseminação criteriosa das informações para as várias audiências.

Ao finalizar o presente artigo, é importante destacar a criação da memória do projeto, atividade que não pode ser desprezada em nenhum momento do seu processo de gerenciamento. As avaliações repetem-se ao longo do tempo e a documentação referente a atividades anteriores pode contribuir para evitar que sejam duplicados trabalhos antes definidos, planejados, desenvolvidos e concretizados. O material para a realização de novas avaliações pode, assim, ser aprimorado à luz de experiências anteriores.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS


HOUSE, E. R.; HOWE, K. R. Deliberative democratic evaluation. In: KELLAGHAN, T.; STUFFLEBEAM, D. L. (Ed.) *International handbook of educational evaluation.* Boston: Kluwer Academic, 2003. v. II. p. 79.

KELLAGHAN, T.; STUFFLEBEAM, D. L. (Ed.) *International Handbook of Educational Evaluation*. Boston: Kluwer Academic, 2003. v. II, section 8, p. 699.

MADAUS, G. *et al*. *Evaluation models*: viewpoints on education and human services evaluation. Boston: Kluwer-Nijhoff, 1993.

MERLEAU-PONTY, M. *Conversas*: 1948. São Paulo: Martins Fontes, 2004. p. 50.

_____. *Merleau-Ponty na Sorbonne*: resumo de cursos, filosofia e linguagem. Campinas: Papirus, 1990.

STAKE, Robert L. Evaluation design, instrumentation, data collection, and analyses of data. In: WORTHEN, Blaine R.; SANDERS, James R. (Ed.) *Educational evaluation*: theory and practice. Worthington, Ohio: Charles A. Jones, 1973.

VIANNA, Heraldo M. *Pesquisa em educação*: a observação. Brasília: Plano, 2000.

# FUNDAMENTOS DE UM PROGRAMA DE AVALIAÇÃO EDUCACIONAL[1]

*Tudo aquilo que sei do mundo, mesmo por ciência, eu o sei a partir de uma visão minha ou de uma experiência do mundo sem a qual os símbolos da ciência não poderiam dizer nada. Todo o universo da ciência é construído sobre o mundo vivido, e se queremos pensar a própria ciência com rigor, apreciar exatamente seu sentido e seu alcance, precisamos primeiramente despertar essa experiência do mundo da qual ela é a expressão segunda. A ciência não tem e não terá jamais o mesmo sentido de ser que o mundo percebido, pela simples razão de que ela é uma determinação ou uma explicação dele.*

*A percepção não é uma ciência do mundo, não é nem mesmo um ato, uma tomada de posição deliberada; ela é o fundo sobre o qual todos os atos se destacam e ela é pressuposta por eles.*

MAURICE MERLEAU-PONTY[2]

**1** Artigo publicado na revista *Estudos em Avaliação Educacional*, n. 28, p. 33-37, jul./dez. 2003.

**2** MERLEAU-PONTY, Maurice. *Fenomenologia da percepção*. 2ª ed. São Paulo: Martins Fontes, 1999. p. 3-6 [Tradução de Carlos Alberto Ribeiro de Moura].

As reflexões sobre avaliação, ora registradas, decorreram de experiências pessoais a partir de 1962 e se expandiram após 1969, compreendendo a publicação de livros e a elaboração de artigos,

especialmente os publicados em *Educação e Seleção* (1980-1989) e em *Estudos em Avaliação Educacional* (1990-2003), ambas edições da Fundação Carlos Chagas, São Paulo, SP. Contribuíram, também, na atualidade, com bastante intensidade, para a configuração dessas percepções, as discussões e o excelente material gerado pelo *Grupo de Trabajo sobre Estándares y Evaluación del Preal, sobre Las políticas de evaluación de logros de aprendizaje en los sistemas educativos de América Latina, no Foro de Discusión 2002*[3], de que participaram educadores da Argentina, Brasil, Chile, Colômbia, Costa Rica, Equador, Estados Unidos, Guatemala, Honduras, México, Nicarágua e Peru . As reflexões aqui consignadas procuram identificar aspectos da ação de avaliar, no conjunto das práticas educacionais, e esperam levar a outros pensares capazes de uma definição dos marcos fundamentais de uma política de avaliação no sistema educacional brasileiro.

**3** Ver: GRUPO DE TRABAJO SOBRE ESTÁNDARES Y EVALUACIÓN DEL PREAL – Foro de Discusión nº 1. "*Las políticas de evaluación de logros de aprendizaje en los sistemas educativos de América Latina*". Síntesis 1, 2, 3 y 4. Resumen Final. GRADE/ REAL. Foro de Discusión 2002. Disponível em: http:// www.grade.org.pe/gtee-preal - Sección Foro.

Os elementos levantados nos vários tipos de avaliação – seja de sala de aula ou de sistemas – devem ser analisados por professores e técnicos especializados nas várias áreas curriculares, a fim de que sejam incorporados ao **planejamento escolar** e contribuam para o processo educacional. A avaliação não é um valor em si e não deve ficar restrita a um simples rito da burocracia educacional, necessita integrar-se ao processo de transformação do ensino/aprendizagem e contribuir, desse modo, ativamente, para o processo de transformação dos educandos.

A expressão "cultura da avaliação" integra, atualmente, a constelação de palavras técnicas no âmbito da comunidade educacional e aos poucos se vai tornado verdadeiro lugar comum, quase que simples figura de retórica; no entanto, é preciso que essa expressão se liberte do seu caráter de mero truísmo e se transforme numa efetiva **política de ação**.

As questões relacionadas ao emprego nem sempre adequado dos instrumentos de medida em avaliação educacional devem ser dimensionadas a fim de que os resultados façam sentido e permitam a orientação das atividades docentes; assim, é importante que se aprofundem estudos ligados à avaliação de processo com o uso de **instrumentos referenciados a critério**, como peça fundamental das atividades de aprendizagem em sala de aula.

Há que pensar em termos de unificação das várias avaliações em relação aos sistemas educacionais; contudo, é fundamental que cada sistema considere a diversidade do seu **espaço**

**social, econômico e cultural**, a fim de evitar interpretações comprometidas e que comparações intra e entre sistemas não levem a colocações destituídas de valor educacional ou que gerem proposições falaciosas.

Os resultados das avaliações não devem ser usados única e exclusivamente para traduzir um certo desempenho escolar. A sua utilização implica servir de forma positiva na definição de **novas políticas públicas, de projetos de implantação e modificação de currículos, de programas de formação continuada dos docentes** e, de maneira decisiva, na definição de elementos para a **tomada de decisões** que visem a provocar um **impacto**, ou seja, mudanças no pensar e no agir dos integrantes do sistema.

A avaliação educacional não subsiste isoladamente, devendo estar associada a outros programas, destacando-se, inicialmente, o de **capacitação docente**, em que a área da avaliação deve integrar, necessariamente, o conjunto das atividades que levam à formação de professores em quaisquer dos níveis de ensino; por outro lado, a avaliação precisa estar ligada à **pesquisa educacional** voltada para a realidade dos problemas educacionais relevantes.

A última década do século XX foi rica de avaliações em larga escala, no âmbito internacional e nacional, neste último caso nos vários níveis da administração governamental; nessa década que se inicia, começo de um novo século, contemplando o passado, devemos nos perguntar: – qual o **impacto dessas avaliações**? E se não houve efetivamente qualquer tipo de impacto, por mínimo que tenha sido, por que não ocorreu? Avaliação e crítica da avaliação (**meta-avaliação**) devem coexistir em um projeto educacional bem estruturado.

A avaliação educacional não objetiva subsidiar, exclusivamente, a **cúpula administrativa**; à avaliação deve seguir-se um trabalho bem planejado de **difusão dos resultados e das suas análises**, a fim de que a sociedade (interna e externa ao sistema) acompanhe o trabalho institucional e possa julgar o seu mérito, inclusive a eficiência transformadora da sua ação.

A partir do espírito de uma nova cultura da avaliação, além da difusão dos resultados, é necessário que se definam diretrizes sobre como usar, produtivamente, esses resultados na melhoria do processo de uma educação que seja eficiente e consequente, evitando-se, desse modo, que os resultados fiquem restritos a uma adjetivação pouco satisfatória.

Uma política de **estruturação de programas** de avaliação não pode ficar restrita ao âmbito da escola, deve, necessariamente, abranger todos os níveis da hierarquia da administração educacional, a partir das Secretarias de Estado, quando for o caso, passando por outros níveis, inclusive técnicos, até chegar à sala de aula e ao professor. A avaliação, consequentemente, não é uma ação isolada, integra toda a comunidade educacional e a própria sociedade.

A definição de uma **política de avaliação educacional** demanda múltiplas considerações, não se restringindo, apenas, ao domínio do conhecimento e ao seu uso na prática. É preciso considerar que, a par do conhecimento para um futuro desempenho, outras dimensões (sociais, culturais e até mesmo éticas) devem ser necessariamente avaliadas e que o programa envolva aspectos quantitativos e qualitativos, incluindo, se possível, interesses, atitudes e valores.

Ao implementar um programa de avaliação há uma preocupação maior em organizar diferentes equipes para fins diversos: administrar, elaborar manuais, construir instrumentos, elaborar questionários, definir logística, orçar despesas, processar dados, analisar informações, elaborar relatórios; contudo, quase sempre se omite a equipe responsável pela **disseminação dos resultados**, junto aos órgãos centrais, às escolas, às famílias, criando-se, assim, um vácuo nas comunicações, talvez o responsável maior pela ausência de um efetivo impacto transformador.

**Avaliações internas** são realizadas pelas unidades do sistema com frequência às vezes modesta, e **avaliações externas** são promovidas por diferentes órgãos oficiais, muitas vezes com a colaboração de instituições privadas, havendo, entretanto, uma falta de sincronia entre essas avaliações, que não têm uma ação efetiva na melhoria da educação. Após sua aplicação, deixam as avaliações de utilizar o seu "potencial energético", entregando-se a uma verdadeira exaustão, até que novo programa se realize, igualmente sem maiores repercussões.

A avaliação não é uma atividade em abstrato, que se realize, como muitas vezes ocorre na prática, ignorando a diversidade dos currículos e a multiplicidade de metodologias de ensino empregadas por professores com diferentes formações (ou ausência de qualquer formação pedagógica), além de posicionamentos diversos quanto às suas áreas de atuação.

É importante que as avaliações sejam discutidas por diferentes segmentos sociais e os seus resultados examinados em função da diversidade das características sociais e em relação à proposta política que define as linhas mestras da educação. A ausência dessas preocupações pode comprometer a continuidade dos programas de avaliação.

Um dos problemas a considerar em um programa de avaliação centra-se na **capacitação técnica** daqueles que se propõem a concretizar o empreendimento. Os "avaliadores" nem sempre dispõem de uma formação especifica, abrangente da complexidade dos diferentes procedimentos avaliativos; executando, desse modo, as suas atividades de maneira amadorística e na base de uma possível experiência pessoal. É o fazer por imitação ou o fazer pela **reprodução** de práticas tradicionais no ambiente escolar. Há, assim, necessidade, talvez urgentíssima, de **formação de quadros técnicos**, a partir de pessoas com experiência docente, para que as avaliações tenham prosseguimento e não fiquem restritas a uma existência episódica sem maiores consequências.

Há que pensar nos projetos de avaliação para o ensino básico no que dispõem as **Propostas Curriculares Nacionais** (PCN), em termos da realidade nacional vivenciada pelos professores. Houve um grande esforço do governo federal, no caso específico do Brasil, em definir, às vezes com excesso de detalhes, o que se propunha para o ensino fundamental e médio. Entretanto, uma pergunta se apresenta de imediato a quem se proponha a analisar o que vem sendo efetivamente realizado: – **as avaliações estão realmente centradas nas propostas curriculares?** A essa indagação segue-se outra: – **as propostas curriculares estão sendo efetivamente seguidas no país?** As propostas curriculares deveriam ser os referenciais para as avaliações, que definiriam padrões mínimos de desempenho, mas uma terceira pergunta se apresenta: — **não seriam os livros didáticos, na sua diversidade qualitativa, os verdadeiros referenciais não apenas para a avaliação, mas para o próprio ensino?**

As instituições educacionais, nos seus diversos níveis, ao detalharem seus programas indicativos das disciplinas que integram o programa curricular, devem, em função dos objetivos institucionais e as características educacionais, culturais e sociais do seu corpo discente, **definir, operacionalmente, cada um dos conhecimentos associados às habilidades esperadas**, a

fim de que possam caracterizar o nível de capacidade de cada um e promover a aceleração dos que se acham em déficit com os padrões estabelecidos. Ao mesmo tempo, impõe-se dar ciência da situação aos interessados, inclusive à família, para que participem da atividade docente. É importante que a sociedade saiba a que a escola se propõe, em termos de competências educacionais e sociais necessárias, para a concretização da cidadania.

Os **padrões para avaliação** devem ser pontos de referência para toda a população e refletir as necessidades dessa população, independentemente de etnia, nível social e econômico, evitando-se discriminações que possam criar diferentes níveis de cidadãos e acentuar ainda mais as desigualdades que marginalizam e estigmatizam os indivíduos. Todos os seres humanos têm condições de realizar diversos tipos de aprendizagem e estruturar novos comportamentos desejáveis, limitando-se a questão, na realidade ao *timing* de cada um, que varia em função de diferentes contingências, como acentuaram Benjamin Bloom e outros, na definição e estruturação de programas de *mastery learning* (aprendizagem para o domínio).

A **definição de padrões ou parâmetros educacionais** condiciona, certamente, o tipo de avaliação a realizar e as características dos instrumentos a empregar nos diversos momentos do processo de aprendizagem que visa a formação e, simultaneamente, a transformação dos alunos. É preciso ressaltar que esses padrões não se devem revestir de um caráter estático de permanência no tempo; ao contrário, devem ser **revistos periodicamente**, elaborados à luz de experiências, modificados, quando for o caso, e até mesmo **suprimidos** se não mais corresponderem à realidade socioeducacional e não atenderem às exigências e necessidades da sociedade.

Ainda que a **avaliação por critério** deva ser norma geral para o ciclo inicial de formação, entre 7 e 14 anos de idade, e **instrumentos por norma** possam ser usados nas demais fases, inclusive nos cursos de nível superior, ou que se façam combinar em um único instrumento a característica de critério e norma, apesar de seus resultados serem mais complexos de interpretar, queremos crer que, independentemente do aspecto formal dos instrumentos, o importante, nessa fase das considerações ora oferecidas, é chamar a atenção para a necessidade de eliminar o caráter coercitivo/punitivo atribuído à avaliação, que sanciona

alunos, impondo-lhes reprovações nem sempre justificáveis e, às vezes, de forma indireta, solicitando que os estudantes de baixo desempenho, mesmo acima de um possível média teórica, se afastem "espontaneamente" da instituição para não prejudicar o prestígio que esta possa usufruir na sociedade, diante de um possível futuro fracasso do aluno, especialmente no acesso ao ensino superior. O que realmente importa é que a avaliação tenha um efetivo caráter formativo e represente um *plus* que faça diferença para melhor na vida do aluno; contudo, para que isso ocorra, é preciso um passo mais amplo no processo de **formação continuada dos professores**, preparando-os para um agir diverso daquele consagrado pela tradição rotineira.

É necessário que não se superestime a questão da **definição de parâmetros e competências** desejadas. Ao lado disso, e prioritariamente, é imprescindível que se estruture todo um processo de formação continuada dos professores e do corpo administrativo para que ambos recebam o embasamento necessário à concretização satisfatória de uma tarefa que certamente demanda grandes esforços de planejamento. São conhecidas as **deficiências profissionais**, sobretudo numa época de pouca valorização do magistério e do pouco atrativo que ele representa para os mais talentosos. Além do mais, ressaltemos a imperiosidade do preparo de **material didático adequado** a diferentes situações, a fim de superar possíveis desvios ou deficiências de aprendizagem e impedir, assim, que se consolidem situações que mais tarde serão difíceis de reverter.

A avaliação não deve utilizar critérios de classificação das escolas (*ranking*), segundo o desempenho da instituição, para fins de divulgação e conhecimento público das que poderiam ser consideradas como sendo as melhores, em função dos seus resultados. As possíveis e reduzidas vantagens do *ranking* no desenvolvimento de uma nova cultura da avaliação acabam por ser superadas por uma problemática bem mais complexa, que é a geração de uma **competitividade negativa** no interior da instituição. O **insucesso em avaliações** pode resultar de numerosos fatores (sociais, econômicos e até mesmo culturais, como no caso bem conhecido recentemente de escolas na Inglaterra, após a chamada era Thatcher) e não, necessariamente, de razões pedagógicas associadas à provável ineficiência do magistério. O possível insucesso, caso seja institucional, deve ser objeto de

**pesquisa, análise e discussões** dentro da própria instituição, com a participação efetiva e solidária da família, que também integra o processo de avaliação.

Um problema a considerar na implantação de um programa de avaliação educacional centra-se na indagação sobre o que fazer com os resultados obtidos. É preciso considerar, por outro lado, se esses resultados serão realmente compreendidos e absorvidos pelos vários segmentos interessados. Certamente que há necessidade do estabelecimento de relação dialógica entre todos os participantes; por outro lado, deve-se pensar, igualmente, na necessidade da formação de equipes técnicas capazes de analisar os dados, identificar problemas e atentar para as implicações desses mesmos resultados na definição das políticas públicas no campo da educação. A sociedade, por sua vez, deve aperceber-se do **significado da avaliação** e das lições que pode proporcionar para toda a comunidade, mesmo para os segmentos que mantêm frágeis relações com o mundo da educação.

A avaliação educacional em uma instituição ou em um sistema não deve resultar de decisões individuais, mas refletir um consenso em que diferentes atores – professores, administradores, técnicos, alunos e a própria família, como intérprete da sociedade – procuram definir os objetivos e finalidades da avaliação, além de outros pontos de relevância, como o tipo de instrumento a utilizar, a definição de responsabilidades dos construtores de questões/itens, a estruturação de procedimentos logísticos para a sua aplicação, a escolha de um tipo de escore ou nota que faça sentido para o grupo avaliado e para a própria sociedade, além, naturalmente, de definir os parâmetros para a análise dos resultados e estabelecer os grupos responsáveis por sua interpretação; contudo, ainda que tudo isso e outros elementos mais sejam estabelecidos de forma criteriosa, é preciso colocar uma indagação relevante em toda e qualquer avaliação: – **o que fazer com os resultados?** Essa é uma questão com inúmeras implicações, que precisam ser consideradas e amplamente discutidas, a fim de evitar que os dados levantados não sejam condenados ao silêncio de um arquivo morto.

Uma questão que merece ser objeto de reflexão consiste na **relação entre o professor e o processo de avaliação**. Qual o uso que os professores fazem dos resultados das múltiplas avaliações a que seus alunos são expostos durante sucessivos anos letivos?

É necessário sempre pensar na **avaliação no contexto de um processo formativo**: – a avaliação para orientar os procedimentos docentes; a avaliação para sugerir novas estratégias eficientes de ensino que levem a uma aprendizagem que seja relevante para o aluno como pessoa humana; a avaliação como um fator de orientação de todo o processo docente, envolvendo não apenas conhecimentos, mas incluindo o despertar de novos interesses e a formação de valores; a avaliação como uma ponte que une professor e aluno visando a um processo interativo gerador de novas aprendizagens; a avaliação como fator capaz de gerar elementos que facilitem a superação dos problemas curriculares e que muitas vezes decorrem de conflitos entre a realidade da escola e o contexto sociocultural em que a mesma se situa. A avaliação, enfim, deve ser um **diálogo de todo o sistema com a sociedade** e do qual o professor participa, mostrando os resultados do seu trabalho, inclusive reconhecendo possíveis erros, mas, ao mesmo tempo, procurando apresentar novas ideias para que a escola se revele uma instituição criativa que consegue superar os obstáculos da burocracia que muitas vezes a sufoca e envolve todo o sistema.

Existe uma necessidade imperativa de que sejam definidos padrões nacionais e regionais que funcionem como referenciais orientadores para os diferentes tipos de avaliação; entretanto, é necessário que esses padrões ou parâmetros não ignorem o caráter vertiginoso das modificações que os conhecimentos sofrem a fim de que a escola não seja reprodutora de elementos obsoletos. Ainda que a escola muitas vezes seja agência revestida de grande conservadorismo, não pode ignorar as mudanças que ocorrem nas várias dimensões da sociedade e permanecer apegada a padrões rígidos, inclusive quanto a procedimentos avaliativos. Precisam ser geradas **novas formas de avaliar** – o que demandará espírito criativo dos educadores –, combinando **elementos quantitativos e qualitativos**, com maior destaque para esses últimos, mas suplantando a dicotomia a que se sujeitam os avaliadores, que se restringem a instrumentos referenciados a critérios e a normas. Ressaltemos, contudo, que os padrões antes referidos não devem ser obrigatoriamente consensuais, impondo-se que na sua definição sejam consideradas as diversidades sociais, econômicas e culturais.

As avaliações, além das características normais relacionadas a diversos tipos de validade (conteúdo, preditiva e de construto),

devem ter, necessariamente, **validade consequencial**. A expressão pode determinar controvérsias, necessitando, portanto, ser plenamente esclarecida. A validade consequencial não se refere a distinções, prêmios e/ou bônus, e muito menos a *rankings* e menos ainda a comparações. É fundamental que os resultados das **avaliações cheguem aos alunos, aos pais, aos educadores e a toda a comunidade educacional**, não devendo ficar restrita apenas aos *policy-makers* da administração escolar. Os resultados das avaliações têm suas implicações, não podendo ser tratados, assim, como uma contabilidade educacional. A avaliação deve ter, forçosamente, consequências, se pretendemos pensar em termos da consolidação da chamada cultura da avaliação. A consequência a que nos referimos está relacionada a novas formas de pensar e agir, demonstrando, assim, que os resultados de uma avaliação fazem diferença e promovem o crescimento da pessoa como ser humano e membro da sua sociedade. Esta sociedade, por sua vez, não pode ficar distanciada do que ocorre na escola, indiferente à constatação dos resultados apresentados, que devem ser discutidos com racionalidade e definidos os caminhos para uma solução sensata dos problemas que essas avaliações refletem.

Avaliar é um agir que se reveste de complexidade, ainda que quase todos – professores e não-professores – se sintam qualificados para expressar **juízos de valor**, cerne de todo processo avaliativo. Se ficarmos restritos ao campo educacional, área de maior interesse no caso presente, constatamos que quase todos temos os nossos sistemas ideais de avaliação, mas que nem sempre resistem a uma análise crítica mais aprofundada. A tendência observada é que tendemos a reproduzir processos de avaliação que nos foram transmitidos por antigos professores, durante nossa formação acadêmica. Isso pode significar alguns acertos, mas, na maioria das vezes, defrontamo-nos com desacertos. Por que? A resposta muito possivelmente está relacionada à formação dos professores ou mais exatamente à ausência de formação dos educadores no campo da avaliação. As licenciaturas concentram-se mais nos conteúdos substantivos do currículo das disciplinas, aos quais são acrescentadas algumas informações pedagógicas, sendo a avaliação de uma forma bastante simplista identificada apenas com a verificação da aprendizagem. Impõe-se uma **nova orientação do processo de formação** dos professores a fim de atualizar a atual geração de educadores e, ao mesmo tempo, criar condições para

que futuras gerações tenham consciência de que **ensinar, aprender e avaliar** constituem um processo interativo contínuo.

A avaliação não pode ignorar as várias dimensões do contexto escolar, tendo em vista a influência que o mesmo tem, e com destaque especial, na definição das diferentes propostas de avaliação, cujo objetivo maior, no final, e sem que paire qualquer dúvida, centra-se na **melhoria do proceder educacional**. É preciso atentar para o fato de que é nesse momento crítico, e não apenas para o professor ou para o avaliador, que todos os envolvidos na ação educativa fazem diferentes opções sobre como obter informações compreensivas que permitam decisões individualizadas ou outros elementos que possibilitem amplas generalizações sobre os diferentes atores dessa complexa teia que é o ato de ensinar e educar. Além disso, é em função desse contexto em que a avaliação se concretiza que diversas opções precisam ser definidas de uma forma consistente pelo professor/avaliador: – **avaliação por critério ou norma; avaliação formativa ou somativa; avaliação interna ou externa** entre outras questões igualmente possíveis e relevantes para os procedimentos subsequentes da avaliação. Não se pode deixar de levar em conta que os elementos obtidos por intermédio das avaliações devem ser, necessariamente, interpretados em **função do contexto** em que foram levantados, o mesmo ocorrendo com a sua **disseminação** para os diferentes segmentos interessados nesse tipo de conhecimento.

A educação, em razão do momento histórico, social e político, considera seus objetivos e define metas a concretizar, não havendo, destaque-se, um modelo único e geral que sirva a todos os povos e a diferentes culturas. Há um ponto sobre o qual parece haver algum consenso, certo grau de concordância entre educadores dos vários sistemas educacionais: – **a educação não visa a proporcionar apenas conhecimento.** O processo educacional procura formar, também, outros tipos de saberes: **o saber ser, o saber fazer e, especialmente, o saber pensar**, que implica, entre outras dimensões, o compreender, o querer, o imaginar e o sentir, como já acentuava Descartes no distante século XVII[4]. E a esses saberes agregam-se, ainda, **habilidades, interesses, atitudes e, particularmente, valores**. Tudo isso é importante e necessita ser considerado durante a avaliação formativa/contínua. Esse considerar leva-nos a um aspecto da avaliação que é

**4** DESCARTES, René (1596-1650). *Princípios da Filosofia*. Portugal: Porto Editora, 1995. Coleção Filosofia Textos. [Introdução e comentários de Isabel Marcelino. Tradução de Isabel Marcelino e Teresa Marcelino.]

fundamental para todos os seres humanos: – **a autoavaliação**, a capacidade de alunos (e por que não dizer, professores, também) se autoavaliarem, procedimento que leva ao autoconhecimento e demonstra a consciência que o indivíduo tem de si mesmo.

As reflexões anteriormente apresentadas consideram de modo subjacente o contexto brasileiro e a sua experiência, rica em ensinamentos, em decorrência dos inúmeros projetos de avaliação implementados a partir do final dos anos 80 e intensificados nos anos 90, por iniciativa de diferentes áreas administrativas (federais, estaduais e municipais), além de algumas poucas de iniciativa da educação privada. Atualmente, possuímos volumosa **soma de dados sobre escolaridade e variáveis socioeconômicas**; contudo, precisamos começar a pensar na sua destinação, ainda que tardiamente, tendo em vista a multiplicidade de usuários possíveis. Uma coisa é certa, e reiteramos mais uma vez, os elementos coletados não podem ser de uso exclusivo da burocracia educacional. É impositivo que os muitos interessados existentes manifestem suas visões e a compreensão que têm dos mesmos. Nesse sentido, com o objetivo de analisar a qualidade da educação e seus problemas, é necessário que se realizem *workshops*, mesas redondas, palestras, debates e outras atividades mais, com a participação de professores, familiares, alunos e diferentes segmentos sociais para que se aquilatem **o valor e o significado dos seus resultados**, demonstrando, dessa forma, a sua compreensão e identificando, também, os pontos críticos que por ventura se tenham manifestado. A partir das informações coletadas é possível definir **projetos de pesquisa** sobre aspectos relevantes, fugindo, assim, à replicação de outras investigações e de temas já exaustivamente estudados em pesquisas anteriores ou realizadas em outros contextos diversos da nossa problemática educacional. Além de definir projetos, é importante que se discuta a própria **política de avaliação** e se tracem novas políticas, além, também, de pensar a sua **fundamentação teórica e as estruturas** que sustentam essas avaliações. Em todos esses aspectos, é importante que se tente envolver universidades e outros centros de excelência que se ocupam com a questão da qualidade da avaliação.

As avaliações de sistemas levantam um número considerável de informações que nem sempre são tratadas adequadamente. É necessário que se decida a *priori* o que fazer com os dados, so-

bretudo porque, tendo em vista o destino escolhido, a decisão tomada vai influenciar de modo considerável **o planejamento da própria avaliação**. Anteriormente, procuramos dar mais destaque à avaliação formativa, ao desenvolvimento individualizado. A estrutura dessa avaliação não terá as mesmas características de uma avaliação baseada em normas, que consideram o desempenho do conjunto amostral, expresso por estatísticas descritivas. É significativo, por outro lado, que se identifiquem os diversos segmentos da sociedade que utilizarão, com inteligência, conhecimento de causa e bom senso, os elementos informativos da avaliação. Nem sempre os mais interessados dispõem de formação profissional adequada para um trabalho em profundidade e que tenha ressonância na comunidade acadêmica. No caso específico do contexto brasileiro, reconhecemos que há interesse em **divulgar resultados** e, nesse sentido, relatórios técnicos são publicados, ainda que com uma certa demora; por outro lado, as autoridades educacionais, a fim de acelerar o processo de disseminação das informações, utilizam-se dos vários órgãos da mídia visando a fazer com que os dados cheguem aos vários segmentos sociais. Isso, entretanto, não basta, não é suficiente, quando não se promovem **estudos analíticos** que identifiquem pontos positivos do ensino/aprendizagem e as defasagens que se constatam, sendo estas bem mais importantes do que aqueles primeiros.

A ação de avaliar sempre provoca reações, muitas das quais com características negativistas, argumentando que apenas aspectos cognitivos são destacados, sem considerar outros aspectos que, por sua natureza, são, muitas vezes, mais importantes do que o simplesmente aprendido. É preciso não esquecer, contudo, que ao avaliar, implicitamente, também se está avaliando algo mais, representado por habilidades, interesses e valores. A avaliação, particularmente a que é realizada em sala de aula, sob responsabilidade direta do professor, é mais importante, sem dúvida, porque não se restringe a um único instrumento, mas resulta, quase sempre, de muitos outros tipos de fazeres, que englobam elementos qualitativos, incluindo entre essas práticas as **técnicas de observação**. No entanto, a avaliação sistêmica, realizada em grande escala, pressupõe, naturalmente, **procedimentos padronizados**, não para fins de comparação, como julgam muitos, mas para gerar um quadro isonômico que dê a todos as mesmas condições para demonstrar as capa-

cidades de cada um por intermédio dos desempenhos específicos que lhes são solicitados. É forçoso reconhecer, contudo, que os procedimentos de avaliação, por mais bem planejados e refinados que sejam os seus instrumentos, nunca oferecem um quadro completo da realidade do ensinar/aprender, pois nunca se conhece a realidade em toda a sua complexidade, assim como, também, por melhores que sejam os indicadores sociais os mesmos não conseguem refletir, com precisão absoluta, a complexidade do mundo social. Sem a avaliação, entretanto, é impossível **formar percepções do processo educacional e da influência da ação educativa da escola-família-comunidade-aluno e professor**.

É necessária uma definição de vários elementos indispensáveis a uma avaliação que atenda a todos os requisitos técnicos, conforme registro anterior; desse modo, entre essas decisões, todas igualmente prioritárias, cumpre estabelecer se a avaliação será por norma ou por critério. Se for por critério, como seria realmente desejável, temos de imediato um sério problema a solucionar: – **qual seria o ponto de corte a ser definido?** Por outro lado, mais um problema, igualmente importante, deve ser equacionado: – **quais os padrões a serem estabelecidos?** A respeito dos critérios é preciso considerar o fato de que esse estabelecimento parte de dados empíricos; portanto, a *posteriori*, ou, então, a partir da experiência de professores da área e/ou de especialistas no campo da avaliação. É preciso lembrar que, na elaboração de provas referenciadas a critérios, é forçoso abranger **amostras representativas de conteúdos e habilidades**, que, supostamente, deveriam ser desenvolvidas na escola. Ainda relativamente à definição de critérios, estabelecidos a *posteriori*, isso não se constituirá em um grande problema se a metodologia empregada for a da **Teoria da Resposta ao Item** (TRI), conhecida em nosso contexto educacional graças ao seu uso no Sistema de Avaliação do Ensino Básico – Saeb[5]. É bastante conhecido nos meios educacionais que o **problema da fixação de padrões** gera controvérsias, mas é uma situação que deve ser encarada e examinada, apesar da sua complexidade. A definição de padrões seria em âmbito nacional ou seriam definidos diferentes padrões regionais, considerando que, muitas vezes, a avaliação não possui caráter censitário, sendo amostral, mas abrangendo uma geografia sociocultural bastante diferenciada? Apesar da complexidade do problema, seria reco-

**5** Seria interessante ler o trabalho de Gregory J. Cizek – *Introduction Achievement Testing in U.S. Schools*, disponível em:http://edex.s3-us-west-2.amazonaws.com/publication/pdfs/cizek_7.pdf, especialmente a discussão sobre testes referenciados a normas e a critério, para maiores esclarecimentos sobre o assunto.

mendável a definição de padrões de desempenho com suas habilidades em termos nacionais, conforme a amplitude espacial da avaliação. Esses padrões poderiam servir de orientação para professores, especialistas em currículo, administradores e pesquisadores, na definição de seus respectivos planejamentos e, inclusive, no caso de pesquisas sobre aprendizagem e rendimento escolar.

Insistimos, reiteradas vezes, ao longo das presentes considerações, que o documento ora apresentado procura traduzir a nossa percepção sobre o problema da avaliação e seu possível impacto nos sistemas educacionais, refletindo-se, dessa forma, a nossa preocupação com a chamada validade consequencial, o impacto que toda essa sistemática exerce nos sistemas de ensino. É preciso, no trato dessas questões, evitar a implantação de certos parâmetros valorativos: – classificações, bônus para os professores, vantagens para os alunos ou premiações, hierarquização das escolas, entre outros, que, no final, acabam por dicotomizar os sistemas, as escolas e os próprios alunos em duas categorias: os **melhores** e os **piores**. Isso determina a perda do espírito de colaboração que deve existir, estabelecendo-se, em oposição, um espírito competitivo entre sistemas, instituições e alunado. É evidente que esse tipo de "consequência" deve ser evitado e superado, quando existe. O importante é que as propostas de avaliação sejam um reflexo da realidade educacional e que possibilitem o autoconhecimento do sistema e o conhecimento do sistema pela comunidade social, que nele investirá em termos de recursos humanos e materiais. Uma avaliação que tenha validade consequencial pode-se transformar, sem sombra de dúvida, em um **processo de certificação de competência**.

Ao longo deste trabalho, procuramos pensar sobre os diferentes problemas ligados à avaliação e suas possíveis soluções a fim de que, aos poucos, mas de forma simples e clara, pudéssemos traduzir as nossas percepções, conforme registramos no início do trabalho. É perfeitamente aceitável que, quando refletimos sobre o que se passa em nosso entendimento, acabamos por gerar e, também, adquirir novos conhecimentos, conforme a visão de Locke[6]. A nossa percepção, desse modo, resultou de uma operação ativa e refletiu a ação do nosso pensamento, mostrando o entendimento que temos da avaliação e o significado que lhe atribuímos no processo educacional.

**6** LOCKE, John (1632-1704). *Ensaios Acerca do Entendimento Humano* (1690) – Livro II – As Ideias, Capítulo IX. São Paulo: Nova Cultural, 1999. p.79-80.

# AVALIAÇÃO EDUCACIONAL: FORMAÇÃO DO AVALIADOR

# AVALIAÇÃO EDUCACIONAL: PROBLEMAS GERAIS E FORMAÇÃO DO AVALIADOR[1]

## 1. INTRODUÇÃO

A importância da avaliação educacional costuma ser ressaltada por todos os que integram a comunidade acadêmica. Até época recente, entretanto, essa mesma avaliação limitava-se à mensuração do desempenho escolar, ou, então, era concebida segundo um modelo simplista, baseado na apresentação de objetivos comportamentais, construção e aplicação de instrumentos, análise dos resultados e elaboração de um relatório final. A avaliação inclui, sem dúvida, todos esses procedimentos, mas a eles não se circunscreve, porque significaria limitar os seus objetivos, que são bem mais amplos.

A avaliação, como área de investigação científica, transformou-se numa atividade complexa. Inicialmente, todo o seu enfoque centralizava-se no aluno e nos problemas de sua aprendizagem; aos poucos, entretanto, sem se afastar desse interesse, modificou a sua orientação e passou do estudo de indivíduos para o de grupos, e destes para o de programas e materiais instrucionais; na etapa atual, preocupa-se com a avaliação do próprio sistema educacional. Se, primeiramente, o interesse maior estava relacionado com problemas de microavaliação, para usar expressão

**1** Artigo publicado na revista *Educação e Seleção*, n. 5, p. 9-14, jan./jun. 1982.

de Payne (1974); agora, sem abandonar estes últimos, ocupa-se com a investigação de questões de macroavaliação.

Alguns dos fatores que possivelmente contribuíram para esta modificação foram:

a) a tomada de consciência de alguns educadores relativamente a problemas educacionais prioritários – educação pré-escolar, educação de carentes, alfabetização, evasão escolar, treinamento de professores, entre outros –, que exigem pronta solução, mas sobre os quais pouco ou nada se sabe, por deficiência de informações, que somente poderia ser superada por intermédio de pesquisas e da avaliação educacional;
b) a insatisfação dos educadores quanto a currículos, programas, práticas de ensino e material didático, impostos ao contexto educacional sem maiores estudos sobre as peculiaridades do meio e da população que sofre os seus efeitos;
c) a reserva de muitos educadores e administradores quanto à qualidade do ensino em seus diferentes níveis. Algumas explicações teóricas têm sido tentadas; contudo, percebe-se que, para a identificação das prováveis causas desse fenômeno, e estabelecimento de alternativas que possibilitem a teimada de decisões, haveria necessidade de procedimentos sistemáticos, representados por trabalhos de avaliação;
d) a aplicação de grandes investimentos financeiros na área educacional, sem que estudos de avaliação de custos e benefícios comprovem a sua influência na promoção da eficiência do processo instrucional.

Outros fatores possivelmente também concorreram para que a avaliação mereça, atualmente, grande atenção. E imperativo ressaltar que, justamente pela importância da avaliação educacional, nesse novo contexto, se deva desenvolver, também, uma consciência crítica, porque nem tudo o que é apresentado como "avaliação, merece realmente essa caracterização; por outro lado, apesar da avaliação ser vista sob um novo enfoque, isso não significa que todos os seus problemas – que são muitos – estejam completamente solucionados.

A avaliação, em muitos dos seus aspectos, ainda apresenta deficiências (STUFFLEBEAM et al.,1971) como:

I. ausência de uma teoria perfeitamente estruturada e que traduza um consenso entre educadores;
II. inexistência de uma tipologia de informações fundamentais para o processo decisório;
III. insuficiência de instrumentos e planejamentos adequados para a avaliação dos diversos fenômenos educacionais;
IV. falta de um sistema que possibilite a organização, processamento e relatório de informações necessárias à avaliação; e, finalmente,
V. carência de elementos qualificados para a realização das complexas atividades que o processo de avaliação exige.

Apesar desses problemas, a avaliação educacional pode oferecer valiosas contribuições para a concretização de mudanças educacionais, que, na verdade, somente deveriam ser empreendidas quando baseadas em conclusões de investigações perfeitamente estruturadas, implementadas e analisadas.

## 2. AVALIAÇÃO EDUCACIONAL - CONCEITUAÇÃO E DEFINIÇÕES

A avaliação educacional, como atividade científica, somente surge na década de 40, com os trabalhos de Ralph W. Tyler, e desenvolvem-se no período de 1960, graças, sobretudo, às contribuições de Lee J. Cronbach, Michael Scriven e Robert E. Stake, entre outros. As várias posições teóricas desses autores, sobre prioridades em avaliação educacional, concorrem para a formulação de diferentes definições desse campo.

A definição mais divulgada de avaliação é a que identifica esta última com o processo de medida. A disseminação dessa concepção resultou, em parte, da divulgação, nos meios profissionais, de obras de cientistas com formação básica no campo da psicometria, como, por exemplo, Robert L. Thorndike e Robert L. Ebel. O estudo das diferenças individuais, por sua vez, concorreu para gerar a crença, bastante difundida, aliás, de que avaliar em educação é medir os resulta dos do rendimento escolar. Assim, avaliação e medida do rendimento são frequentemente usadas como expressões intercambiáveis, e refletem imprecisões no emprego de palavras medir e avaliar.

Medir é uma operação de quantificação, em que se atribuem valores numéricos, segundo critérios preestabelecidos, a características dos indivíduos, para estabelecer o quanto possuem das mesmas. O índice quantitativo, obtido por intermédio da medida, identifica o status do indivíduo face à característica. Relativamente à avaliação, a medida é um passo inicial, às vezes bastante importante, mas não é condição necessária, e nem suficiente, para que a avaliação se efetue. Eventualmente, a medida pode levar à avaliação, que, entretanto, sós e realiza quando são expressos julgamentos de valor.

Avaliar é determinar o valor de alguma coisa para um determinado fim. A avaliação educacional visa, pois, à coleta de informações para julgar o valor de um programa, produto, procedimento ou objetivo (WORTHEN E SANDERS, 1973); ou ainda, a julgar a utilidade potencial de abordagens alternativas para atingir a determinados propósitos. A avaliação refere-se, assim, a atividades sistemáticas ou formais para o estabelecimento do valor de fenômenos educacionais (POPHAM, 1975), quaisquer que sejam.

A avaliação, para alguns, é um processo assistemático, baseado na opinião de um especialista; é um julgamento emitido por um profissional. São comuns, na área educacional, "avaliações" informais para a tomada de certas decisões. Um livro é adotado em vez de outro; uma metodologia de ensino é empregada em substituição a outra, apenas com base em avaliações assistemáticas e impressionistas. A chamada avaliação, nesses casos, limita-se a uma escolha, com base em percepções, da que seria a melhor alternativa. É, pois, uma simples opção, sem fundamento científico. A avaliação, ao contrário, decorre de um esforço sistemático para definições de critérios, em função dos quais se coletam informações precisas para julgar o valor de cada alternativa apresentada. Avaliar é, assim, emitir um julgamento de valor sobre a característica focalizada, podendo esse valor basear-se, parcial mas não exclusivamente, em dados quantitativos.

A Tyler (1942) coube a difusão da definição de avaliação como um processo de comparação entre os dados do desempenho e os objetivos instrucionais preestabelecidos. Essa definição desfruta de grande aceitação nos meios-educacionais e, com pequenas variações, foi incorporada a alguns modelos teóricos, como, por exemplo, o de Hammond (s.d.) e o de Metfessel e Michael (1967).

Stufflebeam *et al.* (1971) desenvolveram um modelo centralizado na ideia de que a avaliação deve permitir aos administradores a tomada de decisões e, coerentemente, definiram avaliação como o processo de identificar e coletar informações que permitam decidir entre várias alternativas.

Outros teóricos da avaliação educacional, juntamente com as suas propostas de estratégias para a investigação avaliativa, propuseram, também, definições que, em maior ou menor grau, são aceitas por muitos praticantes da avaliação educacional. Entre essas definições destacam-se a de Provus (1971), que apresenta a avaliação como um processo de comparação entre desempenho e padrões, e a de Stake (1967), que a caracteriza como descrição e julgamento de programas educacionais.

A avaliação, como campo emergente na área educacional, tem recebido contribuições provenientes de várias fontes, entre as quais se destacam as de Michael Scriven, que marcaram, profundamente, a teoria da avaliação educacional Scriven (1967) concebe a avaliação como um levantamento sistemático de informações e sua posterior análise para fins de determinar o valor de um fenômeno educacional. Essa definição, centralizada no problema do valor, influenciou o pensamento de grande parte dos teóricos e praticantes da avaliação educacional moderna, inclusive de alguns elementos que não se preocuparam em detalhar e explicitar a questão, como foi o caso de Stufflebeam. Analisada a conceituação estabelecida por este teórico, verifica-se que também incluiu um julgamento de valor, ainda que não o tenha explicitado, isso porque escolher essa ou aquela alternativa, isto é, decidir, conforme estabeleceu Stufflebeam, é julgar o valor de uma ou de outra alternativa, optando pela melhor, o que mostra que, na definição de Stufflebeam *et al.*(1971), está implícito, também, um julgamento de valor.

## 3. PESQUISA E AVALIAÇÃO EDUCACIONAIS

Pesquisa e avaliação educacional, apesar de representarem atividades diversas, são frequentemente confundidas. Existem, evidentemente, pontos de contato entre ambas – são formas de investigação científica, usam instrumentos de medida, analisam os dados sistematicamente, às vezes com o emprego

das mesmas técnicas, e muitos dos seus métodos de ação apresentam grandes semelhanças –, o que toma difícil a categorização de atividades que sejam exclusivamente de um único tipo contudo, possuem, também, algumas diferenças substanciais, que merecem ser destacadas.

**1º)** A pesquisa visa a produzir novos conhecimentos, que representem verdades importantes para a ciência, independentemente de sua aplicação prática. A avaliação procura julgar o valor ou a utilidade de um fenômeno, a fim de tomar decisões a seu respeito.

**2º)** A pesquisa, com base na verificação empírica, procura a verdade científica, sem julgar o mérito das relações entre variáveis; ao contrário, a avaliação julga a qualidade de uma determinada relação entre variáveis.

**3º)** A pesquisa procura relacionar variáveis ou fenômenos para estabelecer leis;a avaliação, baseando-se numa escala de valores, descreve um fenômeno específico.

**4º)** A pesquisa, utilizando métodos empíricos ou outros métodos, examina as relações entre variáveis pata obter um conhecimento que seja generalizável, isto é, tenta generalizar os seus resultados para situações comparáveis e quanto maior essa possibilidade, maior o seu significado cientifico. A avaliação concentra-se num fenômeno especifico, procura estabelecer as relações entre as variáveis mais relevantes, e não se preocupa com a possibilidade de generalizar os resultados para outras situações.

**5º)** A pesquisa e a avaliação buscam o conhecimento para melhor compreender os fenômenos educacionais, mas usam esse conhecimento para fins diferentes. A pesquisa, partindo das informações coletadas, visa a extrair conclusões; a avaliação centraliza o seu interesse em tomar decisões.

## 4. MODELOS EM AVALIAÇÃO EDUCACIONAL

A educação é criticada, muitas vezes com veemência, nos seus fundamentos teóricos, na sua estruturação e na sua prática. As decisões tomadas, por pressão dessas críticas, resultam de posicionamentos nem sempre decorrentes de investigações

empíricas, e geram novas críticas e insatisfações, que atingem a validade do próprio sistema educacional. A avaliação, nesse contexto muitas vezes caótico, tomou-se imperativa e exigiu metodologia que possibilitassem a coleta de informações para decisões fundamentadas.

Um esquema de planejamento frequentemente encontrado em projetos de avaliação é o baseado na análise das diferenças apresentada antes e após o tratamento instrucional. Esta estratégia, ainda que útil em certas condições, nem sempre fornece informações detalhadas que permitam tomar decisões complexas. Outra opção estratégica, também amplamente utilizada em avaliação, é a do planejamento experimental, que caracteriza a pesquisa empírica, mas que nem sempre é suficientemente eficiente para a avaliação de alguns fenômenos educacionais, tendo em vista a circunstância de que a avaliação se processa num quadro natural, em que as situações nem sempre são bem estruturadas e, por isso, tomam-se difíceis as condições de controle, exigidas pelo planejamento experimental. Assim, tendo em vista esta problemática, vários especialistas procuraram desenvolver novas estratégias para dar à avaliação um sentido mais eficaz.

Os modelos teóricos apresentados para a solução dos problemas de avaliação educacional variam grandemente entre si, destacando-se como os mais representativos os seguintes:


1. ALKIN, M. C. (1969) – *Evaluation theory development.* Education Comment, 2, 1.
2. CRONBACH, L. J. (1963) – Course improvement through evaluation. *Teachers College Record*, 64.
3. HAMMOND, D. L. (s.d.) – *Evaluation at the local level*. Tuckson, Arizona. EPIC Evaluation Center.
4. METFESSEL, N. S. e MICHAEL, W. B. (1967) – A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. E*ducational and Psychological Measurement*, 27.
5. PROVUS, M. M. (1971) – *Discrepancy evaluation*. Berkeley, California. Mc Cutcham Publishers.
6. SCRIVEN, M.(1967) – The methodology of evaluation, in Stake, R. E. (Bd.) – *Curriculum evaluation*. AERA monograph series on evaluation n9 1.Chicago, Rand Mcnally.
7. STAKE, R.E. (1967) – The countenance of educational evaluation. *Teacher College Record*, 68.

8. STUFFLEBEAM. D.L *et al*. (1971) – *Educational Evaluation and decision making*. Itasca, Illinois, F.E. Peacock.
9. TYLER, R. W. (1942) – General statement on evaluation. *Journal of Educational Research*,35.

As diferenças existentes entre os modelos decorrem do fato de estabelecerem prioridades diversas para os problemas de avaliação educacional. Assim, como exemplificação, e sem aprofundar a análise de todos os modelos, anteriormente mencionados, observa-se que Tyler (1942) se concentra na problemática da convergência entre desempenhos e objetivos instrucionais; Stake (1967) baseia-se na análise de variáveis antecedentes, intermediárias (*transactions*) e resultantes; Stufflebeam (1971), através do exame do contexto - entrada (*input*) – processo e produto, visa a obter informações que permitiam a tomada de decisões pelos administradores.

O avaliador educacional, ao selecionar determinado modelo teórico, para desenvolver um projeto, deverá levar em consideração a natureza do problema a investigar, os recursos disponíveis e a sua própria situação pessoal. Os modelos não se propõem a resolver todos os problemas que se apresentem ao avaliador; objetivam, na verdade, permitir que o avaliador dimensione adequadamente os seus projetos, para evitar que deficiências de planejamento invalidem o processo e levem a falsas decisões.

## 6. FUNÇÕES DO AVALIADOR EDUCACIONAL

A avaliação educacional exige a participação de profissionais especialmente treinados, com experiência no trato de diferentes problemas educacionais, e possuidores de capacitação específica para o exercício da função. A avaliação educacional não deve ser tarefa de responsabilidade exclusiva de professores, pois, em geral, no seu treinamento profissional, apenas recebem informações gerais sobre avaliação, as quais, na maioria das vezes, se restringem à tecnologia da construção de instrumentos para a verificação do rendimento escolar.

O avaliador ou o meta-avaliador (avaliador de avaliações) deve ser um indivíduo capaz de realizar um trabalho científico altamente complexo, que pressupõe habilitações especialmente

desenvolvidas. Sem pretender apresentar uma relação exaustiva dessas capacitações, mas apenas considerando as funções do avaliador numa situação real (MILLMAN, 1975; PAYNE, 1974; WORTHEN, 1975), pode-se estabelecer que, para realizar um trabalho consequente, o avaliador deve ser capaz de:

1. especificar informações necessárias para o de senvolvimento de programas de avaliação;
2. localizar, ler e integrar informações existentes na literatura técnica de pesquisa, medidas e avaliação;
3. analisar possíveis implicações de avaliação anteriores relativamente à avaliação que pretender realizar;
4. definir com precisão o objetivo da avaliação;
5. examinar, criticamente, estratégias de avaliação e selecionar a mais adequada para os fins da avaliação;
6. formular hipóteses ou questões a serem verificadas ou respondidas pela avaliação;
7. especificar os dados necessários para verificar as hipóteses formuladas ou responder às questões propostas;
8. desenvolver planejamentos apropriados para a coleta de dados que permitam examinar as hipóteses ou responder às indagações propostas;
9. selecionar amostras representativas da população para a qual os resultados das avaliações serão generalizados;
10. aplicar o planejamento da avaliação e controlar os fatores que poderiam comprometer a sua validade;
11. identificar padrões ou normas para julgar o valor do fenômeno a ser avaliado;
12. transformar objetivos gerais em objetivos operacionais;
13. identificar classe de variáveis para mensurar;
14. estabelecer critérios para selecionar e desenvolver instrumentos de medida;
15. determinar a validade dos instrumentos de medida usados nas avaliações;
16. usar métodos adequados para o levantamento de dados;
17. controlar o desenvolvimento do programa e identificar desvios de planejamento ou de procedimentos específicos;
18. selecionar e aplicar técnicas estatísticas adequadas à análise de dados;
19. descrever o planejamento da avaliação e os procedimentos de análise em termos de processamento de dados, a

fim de utilizar adequadamente o potencial dos equipamentos eletrônicos;

**20.** interpretar e estabelecer conclusões fundamentadas a partir da análise dos dados coletados;

**21.** elaborar relatórios e discutir as implicações dos resultados da avaliação;

**22.** apresentar conclusões com base nos resultados da avaliação;

**23.** proporcionar retroalimentação sobre o desempenho do programa de avaliação para decisões em caso de sua possível modificação futura;

**24.** demonstrar relações interpessoais adequadas ao funcionamento do grupo de avaliação do grupo administrativo do programa;

**25.** administrar recursos humanos e materiais necessários à implementação de programas de avaliação.

O comportamento apresentado no item 24 é por muitos (MILLMAN, 1975) considerado como sendo o mais importante para o êxito de um programa. O trabalho de avaliação resulta, efetivamente, de um esforço conjunto de pessoas lideradas por um especialista; assim sendo, as características de personalidade do avaliador determinam o tipo de relacionamento do grupo e concorrem para o maior ou menor êxito do trabalho, admitindo-se como satisfatórios os demais comportamentos específicos do avaliador.

As complexas funções do avaliador educacional mostram que o mesmo deve possuir grande maturidade e ampla experiência de ensino ou equivalente; por outro lado, exigem também um treinamento profissional aprofundado, com especial ênfase em análise estatística, medidas e psicometria, métodos de pesquisa, e planejamento de experimentos.

## 7. REFERÊNCIAS BIBLIOGRÁFICAS

MILLMAM, Jason. *Selecting educational researchers and evaluators*. ETS. Princepton, New Jersey: ERIC Clearing house on tests, measurement and evaluation, 1975. 15 p. (TM Report, 48).

PAYNE, David A. Toward a characterization of curriculum evaluation. In: PAYNE,

David A. (Ed.). *Curriculum evaluation*. D.C. Heath: Lexington, Mass., 1974.

POPHAM, William J. *Educational evaluation*. Eaglewood Cliffs, New Jersey: Prentice-Hall, 1975. 328 p

WAHLSTROM, M. W.; TRAUB, R. E. *Evaluation du rendement scolaire*: documentation et informations pédagogiques. Paris: Unesco, 1972. (Bulletin lnternational d'Éducation, n. 184).

WORTHEN, Blaine R. Competencies for educational research and evaluation. *Educational Researcher*, n. 4, v. 1, p. 13-16, Jan. 1975.

WORTHEN, Blaine R.; SANDERS, J. R. *Educational evaluation*: theory and practice. Worthington, Ohio: Charles A. Jones, 1973. 372 p.

# AVALIAÇÃO E O AVALIADOR EDUCACIONAL: DEPOIMENTO[1]

*Uma experiência só faz sua interrupção quando está sendo dita.*
*E se não for dita é, por assim dizer, não existente.*
HANNAH ARENDT

O presente trabalho visa a definir nossa posição em relação a alguns problemas de avaliação educacional. Parece-nos impossível uma discussão da totalidade das questões ligadas à avaliação educacional; desse modo, a argumentação será feita a partir de nossa vivência nessa área, especialmente no período que vai de 1962, quando, mais ou menos formalmente, iniciamos atividades educacionais diretamente ligadas à avaliação, ao ano de 1994, em que começamos a consignar nossas reflexões sobre o assunto, com vistas a registrar a compreensão que possuímos dos diversos temas enfocados.

Ao término de nossa formação acadêmica para o magistério (1952), depois de estudos sobre áreas nem sempre relacionados com a realidade do sistema educacional brasileiro, incluindo, também, matérias pedagógicas, tínhamos uma noção bastante

**1** Artigo publicado na revista *Estudos em Avaliação Educacional*, n. 20, p. 183-205, jul./dez. 1999.

restrita do campo da avaliação, limitada, especificamente, ao rendimento escolar. O assunto, apresentado como um simples tópico da didática geral, era transmitido no curto espaço de uma aula de cinquenta minutos, ao ser analisada a que tão da verificação da aprendizagem. Uma discussão bastante esquematizada da avaliação formal e informal era oferecida sem uma análise aprofundada dos seus fundamentos e de suas implicações, limitando-se, assim, a noções sumárias e gerais sobre como proceder em sala de aula, usando reproduções simplificadas de experiências supostamente consagradas pela tradição pedagógica. Após dez anos de atividades de magistério, em diferentes níveis de ensino, vimo-nos diante de um curso formal de avaliação educacional (1962), ao qual chegamos com algumas mudanças em nossos posicionamentos em decorrência do ensaio de Ebel[2] (1951) e, especialmente, dos vários trabalhos da obra coletiva organizada e editada por Lindquist[3] (1951), que nos deram uma visão mais consistente da avaliação, segundo uma perspectiva quantitativa. Essa abordagem se consolidou com um curso em 1962, ministrado a partir do manual de avaliação da Força Aérea Norte-Americana (USAF), que nos colocou em contato com o pensamento psicométrico, que, no nosso caso, se desenvolveria, inicialmente, com base no pensamento de Guilford[4] (1946) e Flanagan[5] (1951), autores de repercussão em diferentes setores da sociedade educacional norte-americana. O curso por nós frequentado não foi, efetivamente, de avaliação educacional, em seu sentido mais amplo, concentrou-se, apenas, em testes e medidas, especialmente na construção de instrumentos, na discussão de tipos de validade e, sobretudo, na questão da fidedignidade, com seus diferentes métodos de cálculo. Fomos, assim, introduzidos nos fundamentos estatísticos das medidas, o que nos levou a procurar novos conhecimentos nessa área por intermédio de Garrete[6] (1962) e, posteriormente, utilizando a obra de Guilford[7] (1965), entre outras, num esforço de autodidatismo, como ocorre com bastante frequência em nosso contexto educacional, particularmente em assuntos relacionados com a avaliação. Observamos que, decorridos mais de 30 anos da realização desse curso, sempre que se fala em curso/seminário/treinamento sobre avaliação o entendimento é que o mesmo incidirá, obrigatoriamente, em tecnologia da construção

**2** EBEL, R.L. Writing the test Item. In: LINDQUIST, E. F. (Ed.) *Educational Measurement*. Washington, D.C.: American Council on Education, 1951.

**3** LINDQUIST, E. F. (Ed.) *Educational Measurement*. Washington, D.C.: American Council on Education, 1951.

**4** GUILFORD, J.P. NES standards for test evaluation. *Educacional and Psychological Measurement*, 6, 4, 1946, p. 432 e segs.

**5** FLANAGAN, J. C. The use of comprehensive rationales in test development. *Educacional and Psychological Measurement*, 11, 1951, p. 151 e segs.

**6** GARRETE, H. E. A *Estatística na Psicologia e na Educação*. Trad. Eva Nick. Rio de Janeiro: Fundo de Cultura, 1946. 2 vol.

**7** GUILFORD. J. P. *Fundamental Statistics in Psychology and Education*. 41 Edition. New York: McGraw-Hill Book Co., 1946.

de questões objetivas e organização de testes, o que reflete uma falsa concepção do que seja avaliação.

A partir desse curso, assumimos a responsabilidade de orientar um programa de avaliação em uma academia militar que ministrava cursos em nível de 2º grau.

Alguma coisa aconteceu: as avaliações, abrangendo todas as áreas do Ensino Médio, passaram a ser mensais e não mais bimestrais, procuramos capacitar professores na construção de itens, introduzimos, ainda que de forma precária, por falta de hard/software, um sistema de análise estatística das questões, segundo a teoria clássica, e os resultados das provas (testes objetivos) passaram a ser apresentados sob a forma de escores padronizados em função do desempenho do grupo, por intermédio de uma escala de estaninos, que foi criada por Flanagan durante a Segunda Guerra Mundial (1939-1945), a partir da curva normal – modelo esse que, hoje reconhecemos, causou enormes malefícios às ciências humanas, inclusive à educação. A metodologia do seu cálculo foi a apresentada por Durost e Prescott[8] (1966) a qual, por sua simplicidade, permitia que alunos e professores situassem os desempenhos de cada sujeito em relação ao grupo total. Isso representava um avanço, ainda que seja discutível o modelo seguido, baseado em uma curva de probabilidade para fenômenos aleatórios e relativos a um grande número de indivíduos.

**8** DUROST, W. N.; PRESCOT, C. A. *Essentials of measurement for teachers*. New York: Harcourt, Brace, and World, 1962.

A Fundação Getúlio Vargas, no Rio de Janeiro, possuía um centro de estudos de avaliação, sob a supervisão de Ruth Schaeffer e a orientação técnica de Nícia Maria Bessa, e nessa época (1965) promoveu a vinda de especialistas de fama mundial para que transmitissem suas experiências em avaliação. Um deles, Frederick B. Davis ministrou curso sobre medidas, do qual participamos que seguiu orientação tipicamente norte-americana: – testes padronizados, seu uso e interpretação dos resultados. Desta vez, entretanto, foi dado destaque a fórmulas punitivas para a tentativa de acerto casual, muitas vezes apresentadas por intermédio de sofisticação matemática, mas que em nada contribuem para a melhoria do processo de avaliação, como hoje em dia é reconhecido. Esse assunto, que parece preocupar a tantas pessoas não inteiramente identificadas com a mensuração educacional, constou de um amplo ensaio escrito pelo próprio Davis[9] para o *Educational Measurement* na edição

**9** DAVIS, F. B. *Educational Measurement*, Washington, DC. American Council on Education, 1971.

de 1951, organizada por Lindquist. Esse tema punição – para a tentativa de acerto casual foi suprimido da edição seguinte responsabilidade de Thorndike[10] (1971) por ser inteiramente irrelevante, na nossa opinião. A presença nessa mesma época, na Fundação Getúlio Vargas, de Anne Anastasi[11], cujo livro básico (1968)12 teria importância em nossa formação teórica, no final de 1969, e o desenvolvimento por Nícia M. Bessa, em meados da década de 60, de um teste inspirado no *lowa Basic Skills*, para ser utilizado no então Estado da Guanabara, foram fundamentais para a nossa compreensão do processo de medida, especialmente no que diz respeito: a complexidade das chamadas habilidades básicas de crianças da Escola Fundamental. Ao mesmo tempo em que adquirimos *expertise* nessa área, no ano de 1965 procuramos socializar esses conhecimentos através de cursos de curta duração ministrados na Secretaria de Estado da Educação do Estado de São Paulo e em diversas instituições do Estado de Minas Gerais, orientando professores na elaboração de instrumentos – objetivos/não-objetivos – para uma avaliação do rendimento escolar que considerasse inclusive suas implicações sociais – repetência/evasão —, evitando, assim, a realização de um trabalho quase sempre precário e muitas vezes destituído de fundamentação teórica. As intensas atividades nessa área, em diferentes regiões, no período de 1965-67, evidenciaram que a situação não diferia da que conhecêramos há 15 anos, ao terminarmos o curso de bacharelado exigido para a prática do magistério.

O período seguinte, abrangendo os anos de 1967 a 1969, foi marcante em virtude das experiências vivenciadas no exterior, especialmente em Universidades norte-americanas e, em menor grau, em instituições francesas. Se já conhecíamos alguns trabalhos fundamentais da bibliografia norte-americana, passamos a observar mais detalhadamente a prática da avaliação em centros universitários. A Universidade de Michigan (Ann Arbor, Mich.), por intermédio do *English Language lnstitute* (EU), em 1967, e sob a orientação de John Upshur, proporcionaria um trabalho de especialização em medidas do domínio do inglês como segunda língua para estrangeiros. A partir do trabalho de Thorndike e Hagen[12] (1961) , iniciamos estudos para uma maior fundamentação estatística dos instrumentos de medida, graças a obra de Thorndike[13] (1949) em que aborda grande variedade de problemas psicomé-

**10** THORNDIKE. R. L. C Ed. *Educational Measurement*, Washington, DC. American Council on Education, 1971.

**11** ANASTASI, A. *Psychological Testing*. Third Edition. New York. The MacMillan Co. 1968.

**12** THORNDIKE, R. L. e HAGEN, E. *Measurement and Evaluation in Psychology and Education*. New York: John Wiley and Sons, 1961.

**13** THORNDIKE, R. L. *Personnel Selection Test and Measurement Techniques*. New York: John Wiley and Sons, Inc., 1949.

tricos ligados à teoria clássica das medidas, especialmente à questão da validade preditiva, assunto raramente considerado em nosso contexto, inclusive no processo de seleção de recursos humanos para a Universidade. A experiência de Ann Arbor, junto ao ELI, serviu para mostrar que é possível dominar conceitos estatísticos básicos sem um envolvimento mais aprofundado da análise matemática; contudo, para a compreensão da moderna teoria dos testes e de sua fundamentação estatística, uma formação quantitativa é realmente indispensável, a fim de compreender os diferentes modelos matemáticos utilizados no estudo das características humanas.

O ano de 1968 foi rico de novas experiências, que decorreram de estágio no *Centre lnternational d'Études Pédagogiques* (Sèvres, Paris) e em um liceu-piloto na cidade de Toulouse, em situações bastante diferenciadas das que foram experimentadas no contexto norte-americano. Ambas as experiências foram, entretanto, igualmente válidas para a nossa formação na área da avaliação, ao longo de um processo que ainda se desenvolve a cada experiência vivenciada, geradora de conhecimentos que sempre se expandem e renovam. Assim, o avaliador, no dia a dia das suas atividades profissionais, vive contínua construção do conhecimento. A experiência de Sèvres foi válida porque possibilitou acesso a outros centros educacionais e permitiu entrar em contato com especialistas ligados à análise quantitativa de características humanas. Sèvres deu-nos a oportunidade de conhecer Miallaret e Pham[14] (1967) e Barbut[15] (1967) que nos levaram a uma reflexão sobre o alto nível do preparo exigido na formação dos professores. Ambos os livros fogem à orientação norte-americana, pois demandam conhecimentos de matemática avançada, como álgebra de matrizes, fazendo, assim, com que essas obras, destinadas a educadores, tenham, no nosso contexto educacional, uma audiência bastante restrita, sobretudo no caso da obra de Barbut, de grande importância para a compreensão do fundamento matemático de certos modelos usados em educação. Os livros desses autores serviram para uma melhor compreensão do significado da linguagem matemática na análise de problemas das ciências humanas e anteciparam, por outro lado, situações difíceis com que nos depararíamos, ao longo dos anos, à medida que novas perspectivas de conhecimento técnico surgiam em nossa vida profissional. A experiência de

**14** MIALLARET, G.; PHAM, D. *Statistique à l'usage des éducateurs*. Paris: Presses Universitaires de France, 1967.

**15** BARBUT, M. *Mathematiques des Sciences Humaines*. 2 vols. Paris: Presses Universitaires de France, 1967.

Toulouse, ao estabelecermos contatos com a cultura pedagógica dos liceus franceses, na parte relativa à avaliação, mostrou a independência acadêmica dos professores, a sua capacidade de decidir sobre o futuro dos seus alunos, a severidade – muitas vezes exagerada – dos seus exames, que podem gerar frustrações, como as decorrentes do célebre "*baccaleauréat*", que foram parcialmente responsáveis pelos acontecimentos de maio de 1968, cujo significado não foi percebido por De Gaulle, então presidente da França, que considerou tudo simplesmente um *chienlit*, mas cujas raízes penetravam na área educacional e em um sistema de avaliação que precisava ser renovado, corno deve ocorrer em toda e qualquer estrutura educacional, segundo a nossa percepção.

Ainda no final de 1968, tivemos uma sequência de novas experiências que serviram para consolidar antigas vivências e proporcionar novas visões no campo da avaliação educacional, a partir da realização do Mestrado na *Michigan State University* (East Lansing, Mich.), sob a orientação de Robert L. Ebel. As disciplinas do major – *Educational Research Methods, Testing and Grading, Problems of Measurement, Quantitative Methods, Standardized Tests, Advanced Quantitative Methods, Principies of Measurement e Psychological Testing* – foram a base para o domínio da parte substantiva da avaliação educacional, segundo uma perspectiva eminentemente quantitativa, e que foi complementada por um *minor* que fundamentou essa prática, por intermédio de outros cursos, e lhe deu significado: *School Learning*, *Growth and Behavior*, *Philosophy of Education* e *Problems of Higher Education*. A experiência de uma universidade norte-americana é marcante e, no nosso caso pessoal, sentimos que, apesar de altamente competitiva e estressante, exerceu um papel formativo e consolidou antigos conhecimentos. Foi importante, neste novo contexto, o papel de Robert L. Ebel, que nos mostrou que avaliação não é apenas análise estatística, a partir de instrumentos construídos segundo os princípios da tecnologia, que oferecem resultados fidedignos, mas uma atividade que envolve seres humanos e pode ter influência sobre seus destinos, no plano da realização pessoal e profissional. Foi a partir desse momento que começamos a nos preocupar mais seriamente com certos conceitos carregados de abstração, mas que constituem o cerne de toda a avaliação: valores, critérios, objetivos, normas, significância prática, entre outros. A própria avaliação pareceu-nos um conceito abstrato, como reconhecem Madaus, Scriven e Stufflebeam[16] (1993).

**16** MADAUS, G. F.; SCRIVEN, M. S.; STUFFLEBEAM, D. F. *Evaluation Models – Viewpoint on Educacional and Human Services Evaluation*. Boston: Kluver-Nijhoff Publishing, 1993.

A universidade norte-americana é *book-oriented*, exigindo um grande volume de leituras em tempo relativamente reduzido. Em alguns casos, a leitura de dois livros semanais era uma situação comum; por outro lado, um número igualmente grande de artigos era incluído entre as *assigned readings*, aspecto que, no nosso caso, merece destaque, por explicar grande parte das influências recebidas durante nossa formação profissional. Em um dos cursos de Robert L. Ebel – *Problems of Measurement* – tivemos oportunidade de ler alguns autores clássicos na área de medidas, como S. S. Stevens, B. O. Baker, R. M. W. Travers e F. M. Lord (Teoria das medidas); E. F. Gardner, R. L. Ebel, F. M. Lord, W. H. Angoff, E. F. Lindquist, R. T. Lenon (Normas); G. F. Kuder e M. W. Richardson, P. J. Rulon, C. J. Hoyt, R. L. Ebel, L. J. Cronbach, E. D. Cureton, P. Horst, F. M. Lord (Fidedignidade); C. I. Mosier, R. L. Ebel, H. O. Gulliksen, L. J. Cronbach e P.E. Mehl, E.D. Cureton, D. T. Campbell e D. W. Fiske, L. Sechrist, P. E. Meehl e A. Rosen, R. B. Cattell (Validade); J. C. Flanagan, A. P. Johnson, W. G. Findley, M. D. Engelhart, M. W. Richardson, O. K. Buros e L. J. Cronbach (Análise de Itens), cujos artigos foram publicados na antologia organizada por Mehrens e Ebel[17]. A excelência deste material consolidou conceitos e deu-nos um embasamento teórico, além, naturalmente, do domínio de um instrumental necessário ao trabalho que desenvolveríamos após nosso regresso ao Brasil, o que realmente ocorreu, especialmente no período de 1970-86.

**17** MEHRENS, A.; EBEL, R. L. *Principles of Educational and Psychological Measurement, a book of selected reading*. Rand MacNally and Co. Chicago, III. 1967.

Ao discutirmos aspectos de nossa formação acadêmica, especialmente em relação a influências teóricas, não poderíamos omitir o curso de W. A. Mehrens, que adotou como texto básico o livro de D. Magnusson, professor na Universidade de Estocolmo[18]. A obra oferece uma visão bastante lúcida da teoria clássica dos testes, além de ser uma excelente revisão da estatística dos testes, em um nível de matematização suportável. Outras leituras foram igualmente exigidas, conforme a tradição universitária norte-americana, obrigando-nos a uma imersão nas obras de J. P. Guilford, G. A. Ferguson, H. Gulliksen, E. E. Ghiselli, R. L. Thorndike, L. J. Cronbach, F. M. Lord, A. Anastasi e Q. N. Nemar. Houve, portanto, um grande entrosamento entre os cursos ministrados por Ebel e Mehrens, contribuindo, assim, tendo em vista a sua natureza quantitativa, para que nos aprofundássemos na psicometria associada à teoria clássica dos instrumentos

**18** MAGNUSSON, D. *Test Theory*. Addison-Wesley Publishing Co. Reading, Mass. 1967.

de medida, segundo nosso planejamento inicial e a expectativa da instituição brasileira a que nos ligáramos (Fundação Carlos Chagas), quando regressássemos do exterior.

Ainda em relação à nossa vivência em uma instituição norte-americana, com auxílio da Fundação Ford, gostaríamos de acentuar um aspecto importante ligado ao trabalho prático, sempre exigido, aos exercícios constantes, muitas vezes diários, e cobrados para correção e comentários, e as exposições orais, seguidas de debates, com avaliações, inclusive a controversa *peer evaluation*, em alguns casos. A realização de trabalhos, apresentados aos professores em blue books (cadernos padronizados para a realização de exercícios, comuns nas escolas americanas, inclusive na *high school* e nos *colleges*), cria, naturalmente, um ambiente de pressão e competitividade, sobretudo considerando que muitas avaliações tinham os resultados expressos *on the curve*, ou seja, em função do desempenho do grupo. Tudo isso concorre para alimentar o caldo de cultura de um dos elementos mais caros à sociedade americana, a valorização do desempenho e o constante provar que, mesmo não sendo o primeiro, a pessoa se situa no top, entre os melhores, os mais bem dotados e sucedidos, revelando, assim, um outro aspecto da vida social e intelectual da própria universidade: – o culto da meritocracia. A experiência, entretanto, foi extremamente válida e a esse ambiente voltamos em diferentes momentos, que, admitimos, foram sempre enriquecedores, quantitativa e qualitativamente, em nossas atividades profissionais.

Após a experiência universitária norte-americana, e sem uma fase de transição adaptativa à nossa cultura, iniciamos atividades na Fundação Carlos Chagas no final de 1969, centradas na seleção para a Universidade e na de recursos humanos qualificados, especialmente para agências governamentais.

Enfrentamos uma situação inédita, reflexo da massificação do ensino e o consequente afluxo de grande número de estudantes às portas da universidade em busca não apenas do saber, mas, sobretudo de uma qualificação profissional. Ao mesmo tempo que procurávamos socializar conhecimentos e a *expertise* desenvolvidos nos Estados Unidos, tínhamos que defender posições em uma controvérsia inteiramente sem sentido: a falsa dicotomia prova discursiva versus prova objetiva.

Procuramos desenvolver competência, inclusive entre profes-

sores universitários, na construção de itens e no planejamento e montagem de provas objetivas, que eram o instrumento adequado para enfrentar uma situação de exame de massa. Ainda que ataques a esse procedimento fossem originários de segmentos conservadores da comunidade acadêmica, havia repercussões na sociedade, que acreditava na argumentação nem sempre consistente das "autoridades". Isso levou-nos a escrever artigos para divulgação na mídia[19] tentando equacionar o problema diante do quadro revelado pela alta relação candidato/vaga no acesso ao ensino de 3º Graus, e a redigir trabalhos de caráter técnico para as instituições que realizavam seus exames por intermédio da Fundação Carlos Chagas e para a própria comunidade acadêmica[20].

As atividades na área psicométrica intensificaram-se a partir do final de 71, com o aumento do número de instituições que passaram a integrar a Fundação Carlos Chagas, exigindo o deslocamento para diversos pontos do território nacional, especialmente no Nordeste e no extremo Sul, a fim de participar de treinamentos e realizar cursos regulares, em nível de graduação (Faculdade de Filosofia, Ciências e Letras de Araraquara, atualmente integrando a Universidade Estadual de São Paulo) e em nível de pós-graduação, na Universidade Federal do Rio de Janeiro, na área de medidas educacionais, o que resultou na publicação de um livro sobre o assunto[21] posteriormente editado, também, em espanhol.

A preocupação maior em Testes em Educação foi com aspectos ligados ao planejamento dos instrumentos de medida, à redação de objetivos instrucionais operacionais, à tecnologia da construção de questões objetivas e discursivas, ao problema da validade (curricular e preditiva), às questões ligadas à fidedignidade dos resultados e à análise de itens. Tudo isso em linguagem acessível a leitores sem maiores experiências com a estatística dos testes. Aos poucos, o problema da medida da capacidade de expressão escrita começou a adquirir uma dimensão maior, inclusive com o apoio do Ministério da Educação, que, a partir de 1975, oficializou a sua utilização nos exames de acesso à Universidade, com vistas a "resguardar" a língua nacional, problema analisado por alguns setores à luz da segurança nacional. Não nos esqueçamos de que, à época, vivamos em pleno regime militar (1964-85).

Ainda que convencidos da eficiência das provas objetivas na seleção de grande número de estudantes passamos a nos preocupar

**19** Acerto casual em prova objetivas. *O GLOBO*, 04/09/73. A Seleção de Candidatos através de Provas Objetivas. *Folha de São Paulo*, 25/12/73. O que a prova de redação realmente mede? *Folha de São Paulo*, 19/10/75.

**20** Emprego e características de provas objetivas. Ciência e Cultura, vol. 22, nº 3, 1970. Os vestibulares refletem toda a problemática da educação. *Mundo Económico*, vol. IV, nº 5, 1971.

**21** *Testes em Educação*. Editora IBRASA. São Paulo. 1973. *Los Testes em La Educación*. Ediciones Universidad de Navarra SA. EUNSA. Pamplona, Espanha. 1983.

com o problema da medida da expressão escrita[22] analisando experiências realizadas no exterior, especialmente no *Educational Testing Service* - ETS, com particular destaque para o trabalho de Godshalk, Swineford e Coffman[23] realizando estudos empíricos sobre a fidedignidade dos corretores e dos resultados, inclusive usando vários critérios, determinando a validade preditiva e concorrente de provas objetivas na medida da expressão escrita, e verificando, entre outros aspectos, as diversas contribuições da pesquisa educacional para a compreensão dessa medida em situação de exame de massa, como é o rito de passagem do "vestibular". Todo o material escrito (artigos e pesquisas) foi posteriormente publicado em forma de livro[24].

Simultaneamente, outras atividades foram igualmente realizadas, como a análise do possível impacto dos testes sobre o sistema educacional brasileiro[25] tendo em vista a afirmação, aliás, não confirmada empiricamente, de que o exame vestibular estaria moldando, negativamente, o sistema de ensino no Brasil. Além disso, preocupava-nos, sobremodo, a falta de formação técnica dos professores em medidas educacionais, motivando um *paper* que foi apresentado em seminário internacional no Rio de Janeiro, em 1978[26].

O problema do acesso ao ensino superior é recorrente na educação brasileira; desse modo, no final da década de 70, voltou a ser discutido e processos alternativos foram propostos. Velhas questões – provas objetivas versus provas dissertativas – continuaram a ser discutidas *ad nauseam*[27] sem nenhuma comprovação, mas apenas com base em opiniões pessoais, idiossincrasias ou vieses político-acadêmicos. O mesmo assunto voltaria a ser discutido em 1986 e, novamente, tornaria a ser discutidas em 1995, sem maiores consequências práticas. Ainda em 1986, analisamos o problema[28] e voltamos ao assunto em um longo ensaio sobre as origens do vestibular (1911), sua história, sua legislação, suas inovações e retrocessos o título do trabalho reflete a nossa posição sobre o assunto (Acesso à Universidade – caminhos da perplexidade), depois do que passamos a nos dedicar quase inteiramente à avaliação educacional: inicialmente, na área do rendimento escolar; mais tarde, em estudos sobre aptidões e, a seguir, na avaliação de programas e sistemas de ensino.

A intensa atividade na área de seleção não nos impediu de refletir sobre outros aspectos da avaliação, especialmente sobre

**22** Medida da Expressão Escrita. Didata, nº 4, 1976. Redação e medida da expressão escrita: algumas contribuições da pesquisa educacional. *Cadernos de Pesquisa*. São Paulo, Fundação Carlos Chagas, nº 19, 1976. Flutuação de julgamentos em provas de redação. *Cadernos de Pesquisa*. São Paulo. Fundação Carlos Chagas, nº 19, 1976. Aplicação de critérios de correção em provas de redação. *Cadernos de Pesquisa*. São Paulo, Fundação Carlos Chagas, nº 26, 1978. Medida da Expressão Escrita e Prova Objetiva: um estudo preliminar de validade. *Cadernos de Pesquisa*. São Paulo. Fundação Carlos Chagas, nº 38, 1981. Comunicação e Expressão no acesso a Universidade: uma experiência diversificada. Educação e Seleção. São Paulo, Fundação Carlos Chagas, nº 4, 1981. Validade de conteúdo de uma prova de Comunicação e Expressão: análise de alguns problemas. Educação e Seleção. São Paulo. Fundação Carlos Chagas, nº 4, 1961. Redação e Medida da Expressão Escrita: algumas contribuições da pesquisa educacional. Educação e Seleção. São Paulo. Fundação Carlos Chagas, nº 6, 1962. Dupla Correção em provas de redação. In: Comunicação e Expressão. IBRASA. São Paulo. 1963. Provas e Testes no Concurso Vestibular. Educação e Seleção. São Paulo. Fundação Carlos Chagas, nº 12, 1965.

**23** GODSHALK, F. L.; SWINEFORD, F.; COFFMAN, W. E. *The measurement of writing hability*. College Entrance Examination Board. New York. 1966.

**24** "Comunicação e Expressão" – problemas teóricos e práticos de avaliação. IBRASA. São Paulo. 1983.

**25** Impacto dos testes sobre os sistemas e objetivos educacionais: a experiência brasileira. *Cadernos de Pesquisa*. São Paulo. Fundação Carlos Chagas, nº 27, 1978, mais tarde editado por Dockrell, W.B. In: *The Impact of Tests in Education*. IAEA. Princeton, New Jersey, 1980.

a avaliação formativa, para usar a feliz expressão cunhada por M. Scriven, conforme veremos na exposição sobre a lógica da avaliação. O grande drama do ensino brasileiro está na reprovação em várias séries do 1º Grau, como demonstrou Sérgio Costa Ribeiro em vários de seus trabalhos, alguns dos quais editados por nós nas revistas *Educação e Seleção* (1980-89) e *Estudos em Avaliação Educacional* (1990), da Fundação Carlos Chagas. As medidas referenciadas a critério, revigoradas a partir dos trabalhos de Glaser (1963) e por influência de Ralph W. Tyler pareceram-nos o caminho adequado para evitar a situação constrangedora da reprovação, tendo em vista que esse tipo de medida exige uma instrução individualizada e uma avaliação formativa. Chegamos a propor o uso de um coeficiente de verificação da sensibilidade ao processo instrucional[29], mas a proposta não teve maior repercussão junto àqueles que poderiam implementar o projeto: os professores, por razões óbvias, especialmente falta de capacitação na área de medidas e em avaliação.

O ano de 1980 foi rico em experiências pessoais, sobretudo pela participação em um programa internacional, coordenado por William Turnbull, ex-presidente do *Educational Testing Service* – ETS, com o envolvimento da *American University of Cairo, The Hong Kong Examinations Authority* e a Fundação Carlos Chagas[30] para construir um instrumento destinado a avaliar a aptidão numérica e verbal de candidatos à pós-graduação no Brasil e no exterior. Seria um instrumento em quatro línguas: português, inglês, árabe e chinês, em versões tecnicamente equivalentes e que se inspirariam nos modelos do SAT (*Scholastic Aptitude Test*) e do GRE (*Graduate Record Examination*), sem que fossem, entretanto, uma simples reprodução desses instrumentos de comprovada validade preditiva. O projeto, infelizmente, após a pré-testagem dos instrumentos, decorridos quase dois anos de intensos trabalhos, entrou em colapso, como decorrência do falecimento de seu coordenador e principal elemento de ligação com as agências financiadoras.

A participação neste projeto internacional e os trabalhos na área de seleção de recursos humanos evidenciaram a falta de elementos com *expertise* suficiente para o desenvolvimento de projetos nessas áreas, levando-nos a abordar o assunto[31] que é de grande complexidade, tendo em vista a inexistência de centros especializados – a avaliação educacional, infelizmente, não é área de habi-

**26** *Development of Technical Competence of Teachers in Educational Measurement* (paper). International Council on Education for Teaching. Rio de Janeiro, 1978.

**27** Processos alternativos de seleção para ingresso no ensino superior. *Cadernos de Pesquisa*. São Paulo. Fundação Carlos Chagas, nº 34, 1980. Acesso Universidade reflexão sobre problemas atuais. Educação e Seleção. São Paulo. Fundação Carlos Chagas, nº 1, 1980.

**28** Acesso a Universidade – Análise de alguns modelos alternativos de seleção. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas, nº 13 1 1986. Acesso à Universidade Caminhos da perplexidade. Educação e Seleção. São Paulo. Fundação Carlos Chagas, nº 14. 1986. Acesso Universidade – um estudo de validade. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas, nº 15, 1987. Acesso à Universidade – uma reflexão ao longo do tempo. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas, nº 18, 1988.

**29** A perspectiva das medidas referenciadas a critério. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas, nº 2, 1980. Medidas referenciadas a critério: – uma introdução. In: *A construção do projeto de ensino e avaliação*. Fundação para o Desenvolvimento da Educação – FDE. São Paulo, 1990.

**30** Seleção para programas de pós-graduação – um projeto transnacional. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas. Nº 2, 1980.

**31** Avaliação educacional – problemas gerais e formação do avaliador. *Educação e Seleção*. São Paulo, Fundação Carlos Chagas, nº 5, 1982. Qualificação técnica e construção de instrumentos de medida educacional. Educação e Seleção. São Paulo. Fundação Carlos Chagas, nº 10, 1984.

litação em nossas universidades – que promovam uma formação específica para a realização de diferentes atividades que pressupõem, além da experiência docente, conhecimentos diversificados, com profundo embasamento estatístico, mesmo para a realização de estudos qualitativos, como ocorre nos grandes centros universitários, como a Universidade de Illinois (Urbana), no programa de pós-graduação dirigido por Robert Stake.

Procuramos divulgar, ainda em 1982, o pensamento de dois personagens fundamentais na evolução teórica e na prática da avaliação educacional: Tyler e Cronbach[32]. O primeiro, na década de 40, lançou os fundamentos da avaliação educacional; Cronbach, preocupou-se com a fundamentação teórica de sua prática. Acreditamos que não teríamos chegado ao ponto em que nos encontramos se não fosse a colaboração desses dois cientistas sociais, complementada, mais tarde, pela contribuição de outros, como Scriven, Stufflebeam, Stake e Guba, nas suas obras bastante diversificadas.

Ao considerar o período de 1983-84, constatamos que nossas preocupações se diferenciaram consideravelmente, envolvendo problemas psicométricos relacionados com a validade de construto em testes educacionais, seguindo, assim, a linha de Lee J. Cronbach, e com a validade de critério[33]. Nesse ano de 1983, na Pontifícia Universidade Católica do Rio de janeiro (PUC-RJ), realizou-se um seminário que nos permitiu um contato maior com Robert E. Stake, que nos autorizou a tradução de dois de seus *papers*: um, sobre estudo de caso; outro, a respeito de problemas epistemológicos na pesquisa qualitativa/naturalista. Assim, convivemos com o quantitativo e o qualitativo sem maiores traumas, pois julgamos ser inteiramente falsa essa dicotomia que pretende opor uma à outra[34].

A partir de 1986, começamos a nos envolver mais diretamente com o problema da avaliação do rendimento escolar nas escolas de 1º e 2º graus, desenvolvendo projetos financiados pelo Instituto Nacional de Estudos e Pesquisas Educacionais – INEP[35] pelo Banco Mundial e pela Secretaria de Estado da Educação do Paraná evidente que uma avaliação abrangendo 69 cidades dispersas pelos vários Estados do Brasil acaba por mexer com a comunidade educacional, sobretudo tendo em vista a falsa concepção do caráter punitivo da avaliação. Algumas reações revelaram preocupação com aspectos técnicos, numa reação típica de cris-

**32** Avaliação educacional – algumas ideias precursoras. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas, nº 6, 1982.

**33** Validade de construto em testes educacionais. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas, nº8, 1983.

**34** STAKE, Robert E. Estudo de caso em pesquisa e avaliação educacional. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas, nº 7, 1983. Robert E Stake. Pesquisa qualitativa/naturalista: – questões epistemológicas. Educação e Seleção. São Paulo. Fundação Carlos Chagas, nº 7, 1983.

**35** Avaliação do Rendimento de Alunos de Escola do 1ºGrau da Rede Pública: uma aplicação experimental em 10 cidades. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas, nº 17,1988. Avaliação do Rendimento dos Alunos de 2º e 4º séries de Escolas Oficiais do Estado do Paraná. *Educação e Seleção*. São Paulo, Fundação Carlos Chagas, nº18, 1988. Avaliação do Rendimento de Alunos de Escolas do 1º grau da Rede Pública: – um estudo em 39 cidades. *Educação e Seleção*. São Paulo. Fundação Carlos Chagas, nº 20, 1989.

tão-novo que se apega à ortodoxia, antes objeto de contestação pelos recém-conversos. Vimo-nos obrigados, assim, a produzir um documento[36] discutindo objetivos, amostragem e tratamento estatístico dos dados, sobretudo a questão de saber quando NÃO usar certa tecnologia, especialmente tendo em vista o real destinatário da avaliação/pesquisa – o professor em sala de aula.

**36** A prática da Avaliação Educacional: – algumas colocações metodológicas. *Cadernos de Pesquisa*, nº 69. São Paulo.

Ainda que não fosse intenção aumentar o nosso espaço de atividades na área da avaliação, envolvemo-nos, por breve momento, com a avaliação institucional - área pouco desenvolvida entre nós, apesar da importância dos trabalhos realizados pela Universidade Nacional de Brasília (lsaura Belloni) e pela Pontifícia Universidade Católica de São Paulo (Ana Maria Saul), entre outros. A discussão girou, sobretudo, a partir de um texto por nós elaborado sobre a quem caberia a responsabilidade dessa avaliação[37].

A participação em projetos internacionais levou-nos a considerar o problema da qualidade em educação[38] procurando, inclusive, uma forma operacional de medi-la, considerando o contexto em que o processo educacional se desenvolve, as variáveis não diretamente ligadas à escola, mas que afetam a educação, e a ação da escola em termos de entrada, processo e produto. O modelo apresentado procurou demonstrar que o desempenho escolar (e a formação de atitudes) é um dos produtos apenas, não se justificando, assim, a concentração de trabalhos de avaliação unicamente nesse aspecto, como está ocorrendo no Brasil.

**37** Avaliação Institucional: a Universidade (texto proposto para discussão). *Estudos em Avaliação Educacional*, nº 1, São Paulo. Fundação Carlos Chagas, 1990.

**38** Medida da Qualidade em Educação – apresentação de um modelo. *Estudos em Avaliação Educacional*, nº2. São Paulo. Fundação Carlos Chagas. 1990.

Apesar da existência de um número grande de relatórios, há carência de informações sobre aspectos relacionados ao desempenho escolar no final do 2º Grau, ou melhor, os dados coletados nos chamados concursos vestibulares poderiam preencher essa lacuna, mas não são estudados, repousam no cemitério de dados dos arquivos institucionais, ou, o que é mais grave, no arquivo morto de órgãos oficiais. A necessidade de informações urgentes, por solicitação do Ministério da Educação (MEC) – Secretaria de Ensino do 2° Grau, com apoio financeiro do Banco Mundial, fez com que nos envolvêssemos na avaliação de alunos de séries terminais[39] na rede pública e privada, em quatro grandes capitais, verificando as relações entre rendimento escolar e diferentes variáveis socioeconômicas. A avaliação, entre outros aspectos, mostrou que, quando há recursos humanos qualificados, condições materiais, metodologia adequada, recursos didáticos e interesse, entre outros aspectos, – *accountability* –, é possível um ensino eficien-

**39** Avaliação do Rendimento Escolar de Alunos da 3ª série do 2º Grau – subsídios para uma discussão. *Estudos em Avaliação Educacional*, nº 3. São Paulo. Fundação Carlos Chagas, 1991.

te nas escolas públicas, tendo o estudante condições de realizar uma boa aprendizagem, como foi observado nas escolas técnicas. A avaliação, por outro lado, evidenciou a falácia do mito da excelência da escola privada, mostrando, ao contrário, que a sua suposta qualidade nem sempre é verdadeira, sendo superada em muitos casos pela escola pública, desde que bem orientada[40].

No decorrer de 1991, fizemos uma reanálise das avaliações desenvolvidas e essa releitura permitiu-nos uma longa reflexão sobre os fatores determinantes da reprovação e da evasão nas primeiras séries do Ensino Básico[41]. Ao ouvirmos administradores e professores, positivamos a problemática da repetência – soma de mal-entendidos que leva muitos educadores a não se aperceberem de suas calamitosas implicações, inclusive financeiras, além, naturalmente, das psicológicas e pedagógicas, e dos malefícios que determina. Vivenciamos a experiência durante anos, na década de 80, ao realizarmos estudos e avaliações em mais de 300 escolas, abrangendo dezenas de cidades do País, conforme referência anterior. Estes estudos, analisados segundo uma perspectiva temporal, podem ser considerados o início de outros mais sistemáticos, desenvolvidos em nível estadual, a partir do início dos anos 90.

Ao mesmo tempo em que realizávamos diferentes trabalhos em nossa área de concentração, preocupávamo-nos com o problema da meta-avaliação, procurando fazer uma análise crítica[42] do que de mais representativo estava sendo feito no país, chegando a uma conclusão de certa forma pessimista, porquanto, em linhas gerais, situamos a avaliação em nosso contexto educacional na fase da pré-história, por sua preocupação com problemas nem sempre relevantes, limitada a aspectos tópicos, sem maior aprofundamento das questões que interessam, efetivamente, aqueles que militam – os professores – , e que deveriam ser os principais destinatários não apenas das avaliações, mas, também, das pesquisas que se realizam na área da educação, pelo menos em termos teóricos. Ambas, quase sempre, destinam-se a agências financiadoras, atendendo muitas vezes a exigências meramente burocráticas. É a avaliação pela avaliação, a pesquisa pela pesquisa, sem maiores consequências práticas, provocadoras de mudanças no sistema de ensino, nas práticas instrucionais, na elaboração de currículos e na orientação do processo educacional, salvo, naturalmente,

**40** Ver, também, para uma discussão mais ampla desse ponto, o artigo: Avaliação do Rendimento de Alunos de Escolas de 1º Grau da Rede Privada – Pontos Críticos e Convergências. *Estudos em Avaliação Educacional*, nº 7. São Paulo, Fundação Carlos Chagas, 1993.

**41** Evasão, repetência e rendimento escolar – a realidade do sistema educacional brasileiro. *Estudos em Avaliação Educacional*, nº 4. São Paulo. Fundação Carlos Chagas, 1991.

**42** Avaliando a avaliação: - da prática à pesquisa. *Estudos em Avaliação Educacional*, nº 5, São Paulo. Fundação Carlos Chagas, 1992.

as exceções habituais, como é de praxe afirmar.

Iniciamos, a partir de 1991, junto à Secretaria de Estado da Educação de Minas Gerais, um amplo programa de avaliação do sistema estadual de ensino[43] no contexto de um conjunto de atividades ligadas a um programa de qualidade do ensino, parcialmente financiado pelo Banco Mundial. Após a avaliação censitária do Ciclo Básico de Alfabetização, prosseguimos avaliando outras séries do Ensino Básico (5ª e 8ª) e, depois, a avaliação da 2ª série do Ensino Médio e da Habilitação Magistério (3ª e 4ª séries). A avaliação, repetimos, estava inserida num conjunto de outras medidas ligadas à autonomia administrativa, financeira e pedagógica. Uma avaliação desse tipo somente faz sentido se objetiva mexer efetivamente com o sistema, sua administração e, especialmente, com a sua pedagogia, implicando alterações curriculares, a partir da identificação de pontos críticos, além de medidas efetivas para a qualificação dos professores ligados ao ensino das primeiras séries. Apesar da expansão dos programas de avaliação em todo o sistema brasileiro de ensino, nos seus vários níveis – não será mais um modismo imposto pelas agências financiadoras? –, o problema prioritário, a nosso ver, centra-se, realmente, na qualificação de professores, especialmente para o 1º Grau; depois, então, a avaliação, em diferentes áreas visando a aspectos diversos.

A revista Estudos em Avaliação Educacional, em seus números 6 (1992) e 9 (1994), dá uma ideia da complexidade da avaliação de sistemas de ensino, problema que nos levou a delinear uma metodologia[44] considerando aspectos técnicos e oferecendo orientações práticas para a sua concretização, partindo do modelo inicial que desenvolvemos para usar em um programa de medida da qualidade da educação[45].

A experiência desta avaliação demonstrou que o trabalho baseado em população, especialmente, e não por amostragem, somente é possível se houver a colaboração total dos professores identificado com os objetivos do trabalho, a aceitação do corpo discente, consciente da importância do trabalho para a melhoria da sua aprendizagem, e o envolvimento efetivo dos pais em todas as fases do processo, inclusive na análise dos dados e na elaboração dos relatórios finais. O modelo proposto para Minas Gerais foi, posteriormente, com as necessárias adaptações, utilizado em São Paulo e no Paraná, entre outros Estados.

Ao fazermos a revisão dos trabalhos publicados pela revista

**43** Avaliação do Ciclo Básico de Alfabetização em Minas Gerais Estudos em *Estudos em Avaliação Educacional*, nº5, São Paulo, Fundação Carlos Chagas, 1992. Desempenho dos Alunos do CBA em Minas Gerais: análise dos resultados e identificação de pontos críticos. *Estudos em Avaliação Educacional*, nº 6, São Paulo, Fundação Carlos Chagas, 1992. Os alunos da 8ª série do Ensino Fundamental em Minas Gerais: desempenho em redação (análise quantitativa). *Estudos em Avaliação Educacional*, nº 9. São Paulo. Fundação Carlos Chagas, 1994. Atitude em relação à Ciência. *Estudos em Avaliação Educacional*, nº 10. São Paulo, Fundação Carlos Chagas, 1994.

**44** Desenvolvimento de um programa de avaliação do Sistema Estadual de Ensino: o exemplo de Minas Gerais. *Estudos em Avaliação Educacional*, nº 8, São Paulo, Fundação Carlos Chagas, 1993.

**45** Ver *Estudos em Avaliação Educacional*, nº 2, 1990.

*Cadernos de Pesquisa*, ao longo de um período de 20 anos (1972-92)[46], verificamos que há um interesse geral sobre o assunto, especialmente em relação à avaliação do rendimento escolar, e grande preocupação metodológica na abordagem dos diferentes problemas, no entanto, observamos, também, que somente a partir de 1973 as questões ligadas avaliação passaram a merecer um maior enfoque teórico e um interesse maior por metodologias qualitativas, sem, entretanto, um maior conhecimento dos fundamentos dessas mesmas metodologias, grosso modo. Tudo indica que já possuímos um material substancial a respeito da educação no Brasil e que é chegado o momento de uma ação direta, com vistas à alteração do presente panorama, que se revela bastante caótico.

**46** Avaliação educacional nos Cadernos de Pesquisa. *Cadernos de Pesquisa*, nº 89. São Paulo. Fundação Carlos Chagas, fevereiro de 1992.

Ao longo do tempo, procuramos seguir uma linha de coerência em relação ao que pensávamos e fazíamos, sem, entretanto, nos apegarmos a uma rígida ortodoxia, que, geralmente, conduz a caminhos pouco férteis. A presente revisão do nosso pensamento e do trabalho concretizado após 1962 levou-nos a consultar antigas anotações e livros de autores que de uma forma ou de outra contribuíram para a nossa formação. Encontramos na contracapa de um dos livros de Guilford[47], autor que fortemente nos influenciou, uma citação por nós manuscrita em 1967, extraída de obra que infelizmente não anotamos, e que reflete o pensamento positivista de William Thomson, Lord Kelvin (1824-1907), o grande físico inglês: "I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfatory kind; it may be the beginning of know/edge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be" (1883). A transcrição mostra o seu posicionamento claramente positivista e, admitimos, orientou, em grande parte, o nosso proceder, ainda que nunca tenhamos confundido as áreas bastante próximas da medida e da avaliação, podendo ser aquela – a medida – o início desta última, mas não necessariamente, porquanto outras abordagens são igualmente possíveis para a análise dos problemas da educação.

**47** GUILFORD, J.P. *Fundamental Statistics in Psychology and Education*. Mc Graw-Hill Book Co. Fourth Edition. New York, 1965.

O certo é que procuramos acompanhar a evolução do pensamento docimológico e reconhecemos, com Mislevy (1993)[48] que ao longo dos anos algumas obras tiveram grande influência: *An lntroduction to the Theory of Mental and Social Measurements*, de

**48** FREDERICKSEN, N.; MISLEVY, R. J e BEJAR, I.I. *Test Theory for a New Generation of Tests*. Laurence Erlbaum Associates. Publishers; Hillsdale, New Jersey, 1993.

E. L. Thorndlke (1919); *lnterpretation of Educational Measurements*, de T. L. Kelley (1927); *Psychometric Methods*, de J. P. Guilford (1936); *Probabilistic Models for Some lntelligence and Attainment Tests*, de G. Rasch (1960/80) e, particularmente, a obra capital e verdadeiramente enciclopédica que é a *Statistical Theories of Mental Test Scores*, de F. M. Lord e M. R. Novick (1968), trabalho em que colaborou Allan Birnbaum – *Some latent trait models and their use in inferring an examinee's hability* – Parte 5, caps. 17 e segs. – , apresentando a fundamentação da Teoria da Resposta ao Item (*Item Response Theory*). Esses últimos teóricos contribuíram para dar uma nova dimensão à teoria das medidas educacionais (e psicológicas), ajudaram a compreender a relação que existe entre o nível de habilidade dos sujeitos e o escore obtido em um teste, sobretudo a partir de Rasch e outros mais que aprofundaram essa nova visão, geralmente chamada de teoria moderna dos testes, que complementa, mas não invalida, a teoria clássica, na linha de H. Gulliksen e outros.

A teoria das medidas, ao desenvolver-se, passou a exigir certo nível de sofisticação, que dificulta a compreensão daqueles que não têm, por exemplo, conforme dizem Lord e Novick (1968)[49] "competência moderada em cálculo diferencial e integral, além de familiaridade com a linguagem e a mecânica básica da estatística matemática". Isso tem contribuído para dificultar o trabalho de muitos, entre os quais nos incluímos, e afastado outros que se julgam incapazes de penetrar em um campo reservado a alguns poucos eleitos. A sofisticação de certos teóricos, sobretudo considerando o emprego cada vez mais frequente da estatística bayesiana em pesquisas educacionais e psicológicas, demonstra que a área docimológica está a exigir novos talentos com capacitações diferenciadas para acompanhar o desenvolvimento da Psicometria, o que temos procurado fazer, apesar das nossas limitações pessoais decorrentes de uma formação predominantemente humanista.

**49** LORD, F.M.; NOVICK, M.R. *Statistical theories of mental test scores*. Reading, M. A. Addison – Wesley Publishing Co. 1968.

Acreditamos que um avaliador se constrói ao longo de sua experiência profissional, que, no nosso caso particular, contribuiu para que não tivéssemos uma visão maniqueísta de aceitação irrestrita ao positivismo e rejeição incondicional às alternativas que o condenam: pós-positivismo, teoria crítica e construtivismos[50], que conflitam entre si, ainda que se unam na sua oposição ao empirismo. A partir das nossas vivências pessoais,

**50** CUBA, E. G. *The Paradigm Dialog*. Newbury Park. Sage Publications, California, 1990.

quantitativo e qualitativo, objetivismo e subjetivismo, deixaram de ser polos opostos, irreconciliáveis, que não podem coexistir, mas posicionamentos que se completam no exercício da avaliação educacional, onde não devem existir Ormuz e Arimã, mas um pensamento suficientemente flexível para adequar as várias metodologias à diversidade das situações. Por isso, sentimo-nos gratos a autores como Stake, Guba e Lincoln, Stufflebeam, e até mesmo aos radicais Parlett e Hamilton, entre outros, que nos ajudaram a compreender o problema da avaliação educacional e contribuíram, juntamente com Tyler, Scriven e, particularmente, Cronbach, para a formação do nosso pensamento, preocupado com o mundo das realidades, sem cultivar mitos que deformam a visão do avaliador, acentuando mais ainda os seus vieses, que sempre existem.

# INSTRUMENTOS DE AVALIAÇÃO EDUCACIONAL

# QUALIFICAÇÃO TÉCNICA E CONSTRUÇÃO DE INSTRUMENTOS DE MEDIDA EDUCACIONAL[1]

## 1. A MENSURAÇÃO EDUCACIONAL – PROBLEMA GERAL

A análise dos instrumentos de medida do rendimento escolar, ora empregados em nosso meio educacional, revela diversos níveis de qualidade técnica. Ao lado de alguns poucos que realmente demonstram medir aquilo a que se propõem, existe, infelizmente, um número elevado de instrumentos que apresentam completa carência de requisitos técnicos. O problema da qualidade desses instrumentos é grave, sobretudo em virtude da influência que exercem no processo de aprendizagem.

Os instrumentos de medida, independentemente do seu aspecto formal, mas desde que bem construídos, representam um estímulo para o estudante e um desafio ao seu interesse e à sua curiosidade intelectual. É fato reconhecido que os bons instrumentos de medida exercem una função direcional, pois orientam o examinando sobre o que estudar e, mais importante ainda, sobre como estudar. Entretanto, quando certos instrumentos de medida são analisados observa-se que não orientam, mas sim conduzem o estudante a adotar comportamentos sem grande relevância educacional, ou seja, estimulam a aprendizagem do

**1** Artigo publicado em *Educação e Seleção*, n. 10, p 43-49, jul./dez. 1984.

efêmero e do factual, e, assim, transformam-se num elemento de frustação para o estudante, o qual, contrariamente ao que se poderia acreditar, tem plena consciência de não estar sendo avaliado segundo as suas expectativas (EBEL, 1961).

Se há instrumentos de mensuração que não estimulam, não orientam e não avaliam o progresso do estudante, qual seria, então, a razão de ser dessa anomalia? Este estado de coisas não decorreria do tipo de instrumento construído e, particularmente, do tipo de questão elaborada? A pergunta é ociosa e já foi respondida há mais de meio século através de pesquisas empíricas. Existem boas provas de dissertação, assim como há bons testes objetivos. O problema não se concentra neste ou naquele tipo de prova, mas na ausência do domínio da tecnologia dos instrumentos de medida educacional. Alguns responsáveis pela elaboração de provas e exames simplesmente improvisam, quase sempre de boa fé, e, assim, praticam enganos que geram numerosos problemas na área da mensuração educacional.

## 2. OS INSTRUMENTOS E AS MEDIDAS EDUCACIONAIS – PROBLEMAS ESPECÍFICOS

O exame de alguns problemas relacionados com o processo de mensuração demonstra a falta de formação técnica de muitos construtores de instrumentos de medida. Apenas os problemas mais significativos serão discutidos a seguir.

### 2.1. A MAIORIA DOS JULGAMENTOS SOBRE O RENDIMENTO EDUCACIONAL É SUBJETIVA, AINDA QUE MUITOS AVALIADORES ACREDITEM POSSUIR PADRÕES ABSOLUTOS DE JULGAMENTO.

A carência de fid-edignidade dos julgamentos subjetivos acha-se demonstrada por copiosa literatura de pesquisas em educação. A solução para o problema estaria na realização de um julgamento médio por vários examinadores; no trabalho cooperativo para a construção dos instrumentos de medida e, particularmente, no desenvolvimento da compreensão de que somente através do uso de padrões relativos de julgamento, com base no comportamento do grupo de examinados, que constitui o sistema referencial, é possível um julgamento adequado do desempenho escolar.

### 2.2. O TRABALHO DE ELABORAÇÃO DOS INSTRUMENTOS DE MEDIDA É QUASE SEMPRE REALIZADO ÀS PRESSAS E SOB PRESSÃO.

Há, implicitamente, um outro problema, que foge ao nosso campo de indagação, mas que não pode ser evitado, porque influi no estado atual das medidas educacionais. O construtor de instrumentos de medidas, geralmente professor militante, vê-se obrigado a exercer múltiplas atividades desgastantes, física e emocionalmente, e, assim, desconhece o lazer criativo, que lhe permitiria considerar a problemática do processo de orientar e avaliar a aprendizagem. Desse modo, os instrumentos de medida são elaborados às pressas, em véspera de exame e, às vezes, na própria hora do exame, sendo, portanto, instrumentos defeituosos e de má qualidade.

### 2.3. OS INSTRUMENTOS ORA EMPREGADOS SÃO MAL PLANEJADOS E QUASE NUNCA POSSUEM VALIDADE DE CONTEÚDO, POR NÃO CONSIDERAREM UMA AMOSTRA REPRESENTATIVA DE CONHECIMENTOS E CAPACIDADES RELEVANTES.

Um instrumento de medida, qualquer que seja o seu aspecto formal, precisa ser adequadamente planejado. Um conjunto de 50 itens nem sempre é um teste, assim como uma dezena de perguntas de resposta livre nem sempre constitui uma prova de dissertação. A ausência de planejamento reflete-se em interrogação sensatamente apresentadas pelos examinados: – o que mede esse teste? qual o objetivo desse exame? o que pretende o professor com essas perguntas? Nem sempre é possível responder, com sinceridade e sem subterfúgios, a essas perguntas, que são justas e legítimas.

### 2.4. UM GRANDE NÚMERO DE INSTRUMENTOS DE MEDIDA ENFATIZA O TRIVIAL, O DETALHE IRRELEVANTE, SEM CONSIDERAR CAPACIDADES EDUCACIONALMENTE IMPORTANTES.

O importante, para alguns construtores de instrumentos de mensuração, é, por exemplo, a data do descobrimento da América, por Colombo, e não as características do impacto que esse descobrimento teve sobre a evolução da história do mundo moderno. Saber o nome de quem escreveu a obra *De Revolutionibus*

*Orbium Coelestium* é mais relevante, felizmente para um número reduzido de indivíduos, do que compreender o significado da revolução copernicana da ciência moderna. O conhecimento de elementos factuais e especiosos, entretanto, não é exclusivo de um único tipo de prova, parece ser uma tendência geral no atual estado dos instrumentos de medida.

### 2.5. A ESTRUTURAÇÃO FORMAL INADEQUADA DE MUITOS INSTRUMENTOS DE MEDIDA NÃO PERMITE VERIFICAR CAPACIDADES COMPLEXAS, COMO A DE ANÁLISE.

A deficiência formal é muitas vezes considerada por alguns críticos como sendo apanágio exclusivo dos itens objetivos. O argumento, ainda que sem apoio empírico, serve, frequentemente, para justificar e defender os itens de dissertação. Nada mais errôneo, pois, desde que bem elaborados, ambos os tipos de questão podem medir capacidades complexas. Há, na verdade, um problema técnico na estruturação de itens para a mensuração de capacidades complexas, que, infelizmente, nem todos os construtores de instrumentos conseguem solucionar satisfatoriamente.

A análise crítica de textos de dissertação mostra que as chamadas evidências da capacidade de analisar nada mais são do que exercícios de paráfrases de manuais e livros didáticos, sem nenhuma originalidade e profundidade. Entretanto, é um tipo de comportamento que pode ser verificado através de dissertações, desde que esse comportamento tenha sido desenvolvido durante o processo de aprendizagem. Ainda que bizarra, a situação realmente existe; muitas vezes, procura-se verificar comportamentos para os quais o examinando não recebeu treinamento prévio.

O problema é mais ou menos semelhante com relação aos testes objetivos. A falta de qualificação técnica de um construtor de itens não lhe permitirá elaborar uma unidade de informação que exija demonstração da capacidade de análise. Entretanto, a sua verificação é possível; assim, no caso de interpretação de textos literários ou científicos (DIEDERICH, 1955) pode-se exigir que o examinando demonstre esse comportamento através da capacidade de:

a) identificar objetivos e atitudes do autor do texto;
b) estabelecer a ideia principal do texto;
c) mencionar argumentos que apoiam a ideia principal;

d) assinalar os pressupostos em que se baseiam as ideias do texto;
e) indicar diferentes figuras de retórica;
f) criticar a organização do texto;
g) julgar a importância do texto;
h) avaliar o mérito (literário ou científico) do texto.

### 2.6. OS INSTRUMENTOS DE MEDIDA NEM SEMPRE SÃO CONSTRUÍDOS COM A OBSERVÂNCIA DOS PRINCÍPIOS QUE DEVEM ORIENTAR A SUA CONSTRUÇÃO.

Os defeitos de construção costumam ser mais aparente nos testes objetivos do que nos itens de dissertação. Entretanto, ao contrário da crença geral, um bom item de dissertação é coisa rara, pois é mais difícil de ser construído do que um item objetivo, sobretudo quando se pretende um instrumento de alta qualidade (STANLEY, 1958).

Os itens objetivos, construídos sem a observância de normas técnicas, costumam apresentar os seguintes problemas:

a) desequilíbrio na ênfase relativa das dimensões comportamento-conteúdo, o que reflete ausência de planejamento;
b) número reduzido de questões, o que demonstra despreocupação com os problemas de validade e fidedignidade;
c) distribuição defeituosa do índice estimado de dificuldade, geralmente obedecendo a padrões extremos de facilidade ou de dificuldade;
d) irrelevância dos conhecimentos substantivos, inobservância da tecnologia do item e erros grosseiros de edição;
e) favorecimento a padrões regulares de respostas, em virtude da falta de uma distribuição equilibrada das alternativas corretas;
f) inconsistências gramaticais, falta de homogeneidade e plausibilidade das alternativas distratoras, o que favorece o acerto casual.

Os itens de dissertação também apresentam defeitos, ainda que menos visíveis para os não especialistas, mas que nem por isso deixam de ser graves. Os vícios de construção mais frequentes são os seguintes:

I. proposição imprecisa dos temas ou das perguntas, caracterizada pela ausência de determinantes explicativos

dos comportamentos exigidos, o que demonstra falta de planeja mento;

II. emprego indiscriminado da dissertação para verificar comportamentos que poderiam ser positivados por outros meios, o que traduz desconhecimento das várias técnicas de mensuração;

III. irrelevância dos temas apresentados, que geralmente verificam comportamentos cognitivos simples;

IV. número reduzido de questões o que não permite um exame compreensivo e nem garante a validade do instrumento;

V. ausência de critérios pré-estabelecidos e de padrões fixos de correção, que assegurem a fidedignidade dos escores;

VI. influência de efeitos de halo e consequente contaminação dos escores, o que concorre para acentuar a subjetividade dos julgamentos e diminuir a precisão dos resultados.

### 2.7. A MAIORIA DOS INSTRUMENTOS DE MEDIDA, ORA USADOS NO AMBIENTE ESCOLAR, VISAM À AVALIAÇÃO SOMATIVA E NÃO INFLUEM NA ORIENTAÇÃO DO PROCESSO DE APRENDIZAGEM E NEM INFORMAM COM PRECISÃO SOBRE OS SUPOSTOS GANHOS EDUCACIONAIS.

O emprego de instrumentos de mensuração educacional limita-se, frequentemente, ao cumprimento de determinações administrativas, com a finalidade de "dar uma nota" e, após determinado período, aprovar ou reprovar, sem maiores preocupações docimológicas.

Os instrumentos nunca, ou quase nunca, são usados, por exemplo, no início de um curso ou de uma unidade, para fins de determinar o que o aluno sabe e, sobretudo, o que não sabe ou o comportamento que não possui, aspectos esses que permitiriam ao professor estabelecer um plano de trabalho para orientação da aprendizagem. Não há pré-testes e os exames finais não têm o caráter de pós-testes, evidentemente; desse modo, fica o professor impossibilitado de verificar se houve ganhos reais através do processo intencional da educação. Se os instrumentos de medida fossem adequadamente aplicados, poder-se-ia utilizar técnica apropriada (chi-quadrado) e, no caso, verificar a significância da diferença dos desempenhos e comprovar se a diferença resultou de efetiva modificação do desempenho escolar.

### 2.8. HÁ UM DESCONHECIMENTO GERAL DA INFLUÊNCIA EXERCIDA PELO TAMANHO DO ERRO DE AMOSTRAGEM NOS ESCORES DE UM TESTE.

Um teste objetivo ou uma prova de dissertação é uma amostra de conhecimentos e/ou comportamentos possíveis. Assim, qualquer que seja a forma do instrumento, a sua estruturação e faz com base em uma amostra selecionada segundo critérios fixados pelo examina dor. O uso de amostras decorre de limitações óbvias, porquanto, no decorrer de um único exame, é impossível verificar o domínio de todos os conhecimentos e/ou a posse de todos os comportamentos possíveis.

Qualquer que seja o critério de seleção da amostra, inclusive no caso de uma amostra aleatória, comete-se um erro de amostragem. A magnitude desse erro está associada à não representatividade da amostra. Exames não compreensivos, baseados num número reduzido de questões ou de itens, geralmente possuem um erro de amostragem grande. Um bom instrumento de medida procura atenuar a influência desse erro sobre o desempenho do estudante.

Os únicos erros, entretanto, que parecem preocupar são os que resultam do ajustamento incorreto da chave de correção do somatório de escores ou de notas. Há, assim, uma preocupação maior com erros decorrentes de distrações, que na verdade são enganos e podem ser evitados. Se houve enganos e as somas estão corretas, os resultados são considerados precisos, em que pese a influência do erro de amostragem.

### 2.9. A EFICIÊNCIA DA MAIORIA DOS INSTRUMENTOS DE MEDIDA APLICADOS EM NOSSO MEIO EDUCACIONAL NÃO É VERIFICADA PELA ANÁLISE ESTATÍSTICA.

A média e a variabilidade do grupo (desvio padrão) não são determinadas; o grau de dificuldade e o poder discriminativo (validade) dos itens ou questões não são estabelecidos; outros elementos necessários para a análise do instrumento, como o coeficiente de fidedignidade e o erro padrão de medida, não são calculados; entretanto, apesar de todas essas deficiências técnicas, os instrumentos são aplicados e decisões sobre o futuro dos estudantes serão tomadas, enquanto que a maioria dos críticos se preocupa com aspectos formais e ignora outras implicações mais profundas que resultam da carência tecnológica de alguns instrumentos de medida.

## 3. DIFERENTES MODELOS PARA A MENSURAÇÃO EDUCACIONAL

Afirma-se algumas vezes, mas sem muita convicção, que estaria havendo, no momento, uma proliferação de testes objetivos, sobretudo nas escolas de 1º e 2ª graus. Acreditamos que não constitua malefício o uso de testes em qualquer dos níveis de escolaridade, desde que os instrumentos sejam tecnicamente idôneos, usados com propriedade nos casos indicados e os resultados interpretados por pessoa qualificada.

Analisando-se alguns instrumentos atualmente empregados, observa-se que o principal objetivo é coletar dados para a organização de uma rudimentar e discutível contabilidade do rendimento escolar. Além de não considerarem aspectos relevantes para a mensuração escolar – validade e fidedignidade –, esses instrumentos são inconsequentes, porque não informam ao aluno sobre o seu progresso e não possibilitam ao professor verificar a concretização dos objetivos educacionais. À deficiência na construção dos testes e provas associa-se o desconhecimento de técnicas estatísticas elementares; desse modo, o desempenho individual não é analisado em função do comportamento do grupo (LAIDLAW, 1965).

O estado atual das mensurações educacionais simplesmente demonstra que, na escola, por falta de recursos humanos com treinamento especializado, não estão sendo empregados diferentes modelos de mensuração para fins diversos, com evidentes prejuízos para o aluno, o professor e a educação.

Quais os diferentes modelos que poderiam ser utilizados? Testes de critério, testes de predição e testes combinados (critério e predição).

Os testes de critério poderiam informar até que ponto os objetivos de uma unidade (ou curso) foram realmente alcançados e, consequentemente, possibilitariam ao professor uma orientação segura de unidades subsequentes (ou cursos), sem defasagens no rendimento escolar, pois o teste de critério, através da fixação de um nível mínimo de competência (o critério), procura determinar o domínio pelo aluno de pré-requisitos; assim, o estudante, informado de seu sucesso ou insucesso, tem consciência do seu progresso; o professor, por sua vez, pode organizar programas de recuperação para os que não obtiveram êxito e, por intermédio de métodos e técnicas especiais, levar

o aluno a superar suas deficiências de aprendizagem e a acompanhar, sem maiores problemas, o desenvolvimento de outras unidades ou cursos.

Há necessidade de outro modelo – o teste de predição – a fim de verificar o desempenho relativo do indivíduo, comparando o seu rendimento com o do grupo. A função desse modelo não é a de verificar objetivos e determinar deficiências individuais, mas sim a de coletar informações que possam levar à tomada de decisões, como aprovação e orientação vocacional.

Ambos os testes – critério e predição – se completam por suas informações; contudo, um terceiro modelo poderia ser construído para obter os mesmos elementos que os outros proporcionam isoladamente. Ainda que mais complexo, o teste combinado é, na verdade, um teste de predição sobreposto a um teste de critério.

O teste combinado apresenta, inicialmente, um teste de critério, que inclui todos os objetivos a verificar. É o quadro de referência para a elaboração de um teste mais extenso. Os itens estabelecem o desempenho mínimo aceitável, por isso são fáceis, com um índice 90%, porcentagem esperada de acertos.

A partir dos objetivos fixados, são elaborados novos itens para a verificação de diferentes níveis de desempenho além do mínimo aceitável. Os itens da parte de predição sofrem um aumento crescente de dificuldade, cuja amplitude deve variar entre 20% e 80%, a fim de discriminar os melhores. A correção do teste combinado é feita em dois momentos. Inicialmente, é corrigida a parte relativa ao teste de critério, não havendo escores, mas apenas sucesso ou insucesso em alcançar o mínimo aceitável. Os que não foram bem sucedidos são submetidos a diferentes formas de ensino de recuperação até que consigam atingir o critério. Os bens sucedidos têm a segunda parte do teste (predição) corrigida e são atribuídos escores para fins vários.

Os instrumentos aplicados em nossa escola não se enquadram no modelo critério, ainda que estabeleçam um desempenho mínimo – "a média 5" –, porque o ensino não é orientado por objetivos e os instrumentos não seguem a mesma orientação; por outro lado, quando o critério (!) não é atingido, as possibilidades de recuperação são mínimas, talvez um novo exame – "a segunda época" –, geralmente tão duvidoso quanto os exames anteriores, e a ameaça de uma reprovação pura e simples, com a repetição de novo período letivo, cuja eficiência é discutível.

Os mesmos testes também não podem ser considerados de predição, pois nem sempre possuem validade e quase nunca oferecem resultados fidedignos. O que são? o que medem? o que permitem avaliar? – são questões difíceis de elucidar no momento presente.

## 4. QUALIFICAÇÕES TÉCNICAS DO CONSTRUTOR DE INSTRUMENTOS DE MEDIDA

A verdadeira questão, no atual contexto educacional e no referente à avaliação do rendimento escolar, centraliza-se no fato de que muitos construtores de instrumentos de medida educacional não possuem a necessária formação técnica para o exercício de uma atividade específica que exige determinadas qualificações. Utiliza-se, às vezes, de uma tecnologia sofisticada, mas desconhecem os seus fundamentos teóricos.

As qualificações necessárias para o domínio da construção de instrumentos de medida educacional podem ser desenvolvidas através do:

a) conhecimento das vantagens e das limitações dos atuais instrumentos de medida;
b) conhecimento de critérios para o julgamento da qualidade dos instrumentos e dos meios de obter evidências relacionadas com esses critérios;
c) conhecimento de como planejar um instrumento e elaborar diferentes tipos de itens ou questão;
d) conhecimento de como aplicar eficientemente os instrumentos de medida;
e) conhecimento de como interpretar corretamente os escores e outros elementos quantitativos.

## 5. PROGRAMAS PARA O DESENVOLVIMENTO DE COMPETÊNCIAS TÉCNICAS

O desenvolvimento de competências na área de tecnologia dos instrumentos de medida permitirá garantir a validade do processo de avaliação educacional. Faz-se necessário evoluir da atual fase artesanal e ingressar na fase técnica, em que princípios científicos empiricamente estabelecidos substituem o espírito amadorista.

Um programa para esse fim poderia ser estruturado através:

a) da intensificação dos currículos, na área de medidas educacionais, para a formação de professores, nas Faculdades de Educação;
b) da criação e implementação de serviços de avaliação nas escolas de diferentes níveis, afim de:
   **1.** orientar professores na construção de instrumentos de avaliação;
   **2.** definir objetivos educacionais relevantes e prioritários;
   **3.** organizar programas de avaliação formativa e somativa;
   **4.** determinar a eficiência do ensino e diagnosticar pontos críticos;
   **5.** controlar a qualidade dos instrumentos construídos;
   **6.** interpretar os resultados da ap1ícaça'o dos instrumentos;
   **7.** informar e orientar os estudantes sobre o seu desempenho escolar;
   **8.** oferecer suporte administrativo para a elaboração e a aplicação de instrumentos de medida;
c) da organização de programas especiais de curta duração, nas instituições educacionais, para a discussão de problemas, através de seminários e trabalhos práticos.

## 6. RESUMOS

**1.** Os instrumentos de medida educacional, independentemente do seu aspecto formal, quando bem planejados e construídos, estimulam e orientam a aprendizagem do estudante.

**2.** Qualquer que seja o tipo de instrumento, é necessário o domínio da tecnologia da sua construção, a fim de que sejam meios válidos de mensuração e fidedignos os resultados da sua aplicação.

**3.** O sistema de mensuração ora em prática apresenta problemas que revelam a inobservância dos fundamentos teóricos e dos princípios tecnológicos que orientam a elaboração de instrumentos usados num programa de medidas.

4. Apesar da existência de diferentes modelos para mensuração educacional, os mesmos não estão sendo utilizados em nenhum dos níveis de escolaridade.
5. Através de um treinamento especializado, é possível desenvolver capacitações técnicas a fim de permitir a introdução e a implementação de programas válidos de medidas educacionais.

## 7. REFERÊNCIAS BIBLIOGRÁFICAS


CRONBACH, L. J. *Essentials of psychological testing*. 3th ed. New York: Harper and Row, 1970.

DIEDERICH, P. B. - Making and using tests. *English Journal*. Illinois: NCTE, 1955.

EBEL, R. L. Improving the competence of teachers in educational measurement. *The Clearing House*, New York, v. 36, n. 2, October 1961.

_____. *Essentials of educational measurement*. Englewood Cliffs, N J: Prentice-Hall, 1972.

ENGELHART, M. D. What to look for in a review of an achievement test. *Personnel and Guidance Journal*, n. 42, p. 616-19, 1964.

KATZ, M. *Selecting an achievement test*: principles and procedures. Princeton: Educational Testing Service, 1961.

LAIDLAW, W. J. Teacher-made test: models to serve specific needs. *The Clearing House*, February 1965.

STANLEY, J. C. ABC's of test construction. *National Educators Association Journal*, April 1958.

# NATUREZA DAS MEDIDAS EDUCACIONAIS[1]

## 1. INTRODUÇÃO

A mensuração de variáveis educacionais e seu tratamento quantitativo apresentam inúmeras dificuldades. A falta de definição precisa das variáveis, a frequente impossibilidade de construir instrumentos de mensuração adequados e, particularmente, a divergência quanto ao significado das medidas (LORGE, 1951), entre outros problemas, são elementos que concorrem para a configuração de uma situação complexa na área educacional e bem diferente da que resulta, por exemplo, das mensurações físicas.

## 2. MEDIDA DE ATRIBUTOS

Antes da discussão dos vários níveis de medida, é necessário considerar alguns aspectos específicos: "— que é medir? o que se mede efetivamente? as medidas educacionais são diretas"?

A palavra medida é empregada com diferentes significados e aplicada para os fins mais diversos, podendo traduzir:

- o ato ou processo de determinar a quantidade, duração ou dimensão de uma coisa;

**1** Artigo publicado em *Educação e Seleção*, n. 9, p. 7-16, jan./jun. 1984.

- o instrumento pelo qual o processo é realizado;
- as unidades em que os instrumentos são graduados;
- os resultados do ato de medir (JONES, 1971).

Medir, no seu sentido mais amplo, segundo Stevens (1946), é atribuir números a objetos ou acontecimentos segundo certas regras. Esses números, naturalmente, representam propriedades ou características. A definição de medida, conforme acentua Kerlinger (1973), não faz referência à qualidade dos procedimentos de medida. É importante que se compreenda que as medidas educacionais envolvem, básica e essencialmente, a mesma teoria e os mesmos procedimentos gerais de outros tipos de medidas, como as físicas. Destaque-se, ainda, como enfatiza Kerlinger (1973), que a definição de medida, nos termos apresentados e desde que sejam definidas as regras, possibilita, teoricamente, qualquer mensuração. Outro aspecto importante, que decorre da definição de Stevens (1946), resulta da importância de estabelecer com adequação as regras de medida, sem o que o processo de medida será invalidado.

É pouco provável que se chegue a um consenso sobre o exato significado da palavra medida, considerando-se que o emprego de mensurações é o mais variado possível. Entretanto, quaisquer que sejam as medidas, físicas ou não físicas, incluindo-se entre estas as educacionais, psicológicas, sociológicas etc., elas se referem a atributos, propriedades ou características dos objetos, conforme destaque anterior, e são realizadas para obter informações que possibilitem inferências sobre os objetos (LORGE, 1951; JONES, 1971). Assim, não se mede um estudante, mas a sua capacidade (atributo), com a finalidade de descrever o seu rendimento escolar e prever o seu desempenho subsequente (JONES, 1971). Medir é, portanto, atribuir número a quantidades do atributo dos objetos, segundo determinadas regras. Usando-se um sistema de números, um certo atributo é quantificado, mas o problema do é de fácil solução. Alguns atributos podem ser facilmente medidos e, portanto, quantificados; outros, ao contrário, especialmente os de interesse na área educacional – rendimento escolar, habilidades, aptidões, atitudes etc. –, por não possuírem uma definição operacional precisa, são de difícil mensuração. Além disso, a construção dos instrumentos de medida reveste-se de grande complexidade e nem sempre é realizada de forma inteiramente adequada (JONES, 1971).

## 3. MEDIDA DOS EFEITOS

O interesse por diferentes conjuntos de atributos, na área da docimologia educacional, exige um número variado de procedimentos. Algumas medidas podem ser realizadas direta mente; outras, entretanto, somente são obtidas indiretamente, por seus efeitos, como é o caso das medidas educacionais. Quando se aplica um instrumento para fins de medida do rendimento escolar, pressupõe-se que haja uma correspondência entre os diferentes níveis de desempenho e os diversos níveis de conhecimento, ou seja, infere-se que ocorra uma relação entre o efeito (desempenho no teste) e o atributo mensurado (rendimento escolar).

As medidas, na área das ciências do homem, são, muitas vezes, indiretas, como ocorre no campo da educação, em que são medidas propriedades, características, atributos dos indivíduos. Pode-se dizer, com mais rigor, que, na realidade, se medem elementos indicativos das propriedades dos objetos ou das características dos indivíduos. Essas propriedades, na área educacional, são inferidas a partir da observação de presumíveis indicadores dessas propriedades. Assim, para realizar medidas educacionais, é indispensável o estabelecimento de definições operacionais, que determinem os elementos indicadores dos atributos a serem inferidos, ou seja, é preciso que se estabeleçam construtos (KERLINGER, 1973).

O problema, em qualquer tipo de mensuração, inclusive as educacionais, centra-se na necessidade de especificar e controlar as condições de observação, a fim de que fatores estranhos não interfiram no processo e prejudiquem as inferências. A especificação e o controle de variáveis do comportamento humano, ao contrário do que ocorre com as variáveis físicas, reveste-se de enorme complexidade e, geralmente, a precisão dessas medidas é afetada por um componente – o erro de medida –, que resulta da impossibilidade de controlar todas as condições de observação. Acresce, ainda, o fato de que as observações estão sujeitas à variabilidade humana, fazendo-se necessária a aplicação de instrumentos adequadamente construídos, a fim de que o processo de medida não seja deformado em decorrência da influência do erro.

## 4. NÍVEIS DE MEDIDAS

O processo de medida, conforme discussão no item anterior procura informações sobre os atributos dos objetos. Essas informações decorrem dos numerais atribuídos às características medidas e dependem do tipo de escala utilizada. As escalas mais comumente empregadas apresentam-se em quatro níveis, que, a partir do mais baixo, são: nominal, ordinal, intervalar e de razão.

### 4.1. ESCALA NOMINAL

A escala nominal de medida é a mais limitada, a mais primitiva das escalas, no dizer de Stevens (1946), e sua natureza é apenas classificatória. São fixadas categorias bem definidas e delineadas, cujos elementos têm como propriedade fundamental a equivalência ou igualdade. Todos os elementos na mesma categoria são equivalentes (iguais) relativamente à categoria (atributo) medida (ARMORE, 1967). A relação de equivalência é reflexiva (x = x para todo x), simétrica (se x = y, então y = x) e transitiva (se x = y e y = z, então x = z) (GUILFORD, 1954; SIEGEL, 1975).

A partir de semelhanças e diferenças entre os objetos, todos os que possuem algo em comum são incluídos numa classe ou categoria. Os elementos de uma categoria são equivalentes (iguais) relativamente à característica mensurada e nenhuma outra informação é proporcionada por essa escala, além da equivalência. Suponhamos os seguintes atributos: – sexo, nacionalidade, religião e ocupação. A classificação dos indivíduos, segundo esses atributos, seria: sexo (feminino, masculino), nacionalidade (brasileiro, chinês, italiano etc.), religião (católica, protestante, espírita etc.) e ocupação (carpinteiro, motorista, pedreiro etc.).

As categorias podem receber nomes ou números, mas esses números são apenas rótulos e somente servem para identificar os indivíduos numa classe. Os números são usados sem que se pretenda a realização de qualquer operação matemática; por outro lado, não refletem quantidades do atributo. O atributo sexo foi categorizado em masculino e feminino. Algumas vezes atribui-se 0 (zero) ao sexo feminino e 1 (um) ao masculino. Qualquer outro número também serviria, sem que isso implicasse a modificação da natureza da classe ou categoria. A única propriedade dos números aplicável a esse nível de medida é a da diferença, isto é, quando a medida é nominal, um número é diferente do outro. Se à classe A é atribuído o número 2 e à classe B o número 3, isso

apenas significa que A e B são diferentes quanto ao atributo medido. Se somássemos 2 e 3, referentes às classes A e B, como interpretar os resultados? Além dessa limitação, os números, numa escala nominal, não permitem inferir sobre diferenças na quantidade do atributo medido.

As categorias, na escala nominal, podem ser representadas por qualquer símbolo, além dos números (cores, letras, desenhos etc.). Se considerarmos mais detidamente esses símbolos, veremos que o seu significado é restrito à identificação dos sujeitos, sendo sem sentido a sua aritmetização com outros símbolos, que são simples "etiquetas".

A escala nominal é, assim, o nível mais baixo de mensuração e representa a fase inicial de operações mais complexas (GUILFORD, 1954). Há, entretanto, quem não considere a classificação como uma medida, pois

- o atributo, que serve de base para a classificação, não precisa ser interpretado em termos de grandeza;
- a inclusão de um elemento numa categoria não precisa ser representada por um número (JONES, 1971).

Outros, ao contrário, como é o caso de Kerlinger (1973), acreditam que desde que a definição de medida seja satisfeita e os elementos categorizados possam ser contados e comparados, os procedimentos nominais são uma medida. A própria expressão escala nominal é, também, contestada por alguns, pois a palavra escala dá ideia de um *continuum* que possui a propriedade da ordenação, o que não ocorre nas chamadas escalas nominais. Se, entretanto, dermos ao termo escala o significado de "aquilo que discrimina", o emprego da expressão é legítimo (GUILFORD, 1965). Apesar das restrições que possam ser feitas à escala nominal, sua importância, conforme destaque anterior, é grande, tendo em vista que a categorização constitui a base de todos os tipos de mensuração (GUILFORD, 1965).

### 4.2. ESCALA ORDINAL

A escala ordinal ou escala de postos reflete a posição ou a importância relativa da medida de um atributo. A escala ordinal apresenta duas propriedades: – equivalência e importância relativa (maior do que; menor do que). Sempre que a relação > for válida para todos os pares de classe, a escala é ordinal (SIEGEL, I 975).

A relação maior do que é irreflexiva (não é verdade que, para qualquer x, se tenha x > x), assimétrica (se x > y, então y l > x) e transitiva (se x > y c y > z, então x > z) (GUILFORD, 1954; SIEGEL, 1975). A principal condição a ser satisfeita para que se tenha uma escala ordinal é a da transitividade.

Sempre que um atributo existe em diferentes graus, é possível medi-lo numa escala ordinal. Os objetos são relacionados uns aos outros e ordenados segundo a quantidade do atributo que possuem. A ordem relativa dos grupos classificados representa um nível mais elevado de medida. Enquanto a escala anterior, a nominal corresponde a uma classificação qualitativa, a medida ordinal é uma classificação quantitativa que possibilita comparações entre grandezas. Os números, entretanto, na escala ordinal indicam posições, postos e nada mais; não indicam quantidades absolutas e também que os intervalos entre os números são iguais.

As relações "maior do que", "menor do que" e "igual a", nas escalas ordinais, ocorrem porque se supõe que os indivíduos ocupem uma posição no *continuum* que representa o atributo medido. A posição traduz a quantidade do atributo e informa a direção do *continuum*, o que permite dizer se a posição de um objeto é maior do que, menor do que ou igual à posição de outros objetos (MAGNUSSON, 1967). Usa-se, portanto, a propriedade da ordenação. Se o número atribuído ao objeto. A é maior do que aquele que caracteriza B, isso significa que A possui mais quantidade do atributo do que B. A essência da escala ordinal é, como já se ressaltou, o conceito de "maior do que".

A escala de dureza dos minerais é um exemplo típico de medida em nível ordinal, conforme assinala Stevens (1946). Se o mineral A risca o mineral B, então, aquele é mais duro do que este. Suponhamos que aos minerais A, B, C e D foram atribuídos os números 8, 6, 4 e 2, respectivamente. Sabemos qual o mais duro e qual o menos duro, mas não podemos afirmar que a diferença entre a dureza de A e B é igual à diferença que existe entre a dureza de C e D. Igualmente, nada é possível dizer sobre o número de vezes que um atributo é maior ou menor do que o outro; assim, não faz sentido afirmar que o mineral A é quatro vezes mais duro do que o mineral D, simplesmente porque o número que expressa a dureza de um é quatro vezes maior do que aquele que expressa a dureza de outro mineral.

Exemplificando na área educacional – dez estudantes foram classificados segundo a sua habilidade numérica e receberam posições que variaram de 1 (maior habilidade) a 10 (menor habilidade). O estudante que ocupa a posição 5 tem mais habilidade do que os situados nas posições 7, 8 e 9 e menos habilidade do que os localizados nas posições 4, 3 e 2. A escala ordinal indica as relações de maior do que (>) ou menor do que (<), além da equivalência (=), mas não informa sobre o quanto existe de diferença.

Quando se tem uma escala ordinal, os números não permitem inferências sobre a quantidade da diferença entre um atributo e outro. Admitamos, para fins de argumentação, que um grupo de estudantes foi classificado segundo o atributo habilidade mecânica. O estudante com mais habilidade recebeu a posição 1, outro a posição 2 e assim sucessivamente. Quando se consideram indivíduos em posições adjacentes, suponhamos os de posições 3, 4 e 5, as diferenças em habilidade mecânica podem ser grandes ou pequenas, mas a escala não inf9rma a esse respeito. Por outro lado, as diferenças numa escala ordinal não são necessariamente iguais.

Além desses problemas, as posições numa escala ordinal apresentam outro inconveniente: elas não são fixas, modificam-se e a sequência dos números se altera quando o número de indivíduos do grupo observado se modifica. Um grupo de 30 estudantes é classificado segundo o atributo altura. Se a esse grupo forem acrescentados outros estudantes, as posições ocupadas pelos indivíduos do grupo inicial muito possivelmente se modificaria.

A ordenação é, pois, o aspecto fundamental da escala ordinal, que, entretanto, apresenta limitações matemáticas. Não faz sentido empregar as operações aritméticas comuns aos números ordinais. Assim, se numa distribuição de postos somarmos o primeiro ao quinto sujeito, o resultado obtido não autoriza qualquer comparação com o sexto colocado nessa ordenação.

A escala ordinal apresenta, ainda, duas outras deficiências que têm implicações nas medidas educacionais. Primeiramente, as medidas ordinais não informam o desempenho dos elementos como um todo. Suponhamos que quatro indivíduos foram ordenados segundo um determinado atributo e que não houve empates nessa ordenação. Assim, temos os postos 1, 2, 3 e 4, que não dizem se o grupo, no seu conjunto, é excelente,

bom ou medíocre, relativamente ao seu desempenho. Outra deficiência das escalas ordinais refere-se ao fato de não proporcionarem informações sobre a dispersão dos desempenhos, ou seja, a partir das posições é impossível estabelecer a diferença entre os vários postos e saber se a dispersão entre eles, ou comparativamente a um outro grupo, é grande ou pequena.

Os escores de testes de rendimento escolar são expressos em escalas ordinais e por isso possuem grandes limitações. Admitamos que um grupo de 20 estudantes foi submetido a um teste de escolaridade. Se esses alunos forem ordenados quanto ao desempenho no teste, teremos as posições 1, 2, 3, .... 18, 19 e 20. Se tivermos outro grupo e o submetermos ao mesmo teste e ordenarmos os resultados da mesma forma, teremos posições semelhantes às primeiras. Quaisquer que sejam os grupos, as posições serão sempre as mesmas e nada informarão sobre as características do grupo e dos indivíduos, que podem ocupar as mesmas posições apesar de serem diferentes, ou as diferenças entre posições adjacentes podem ser grandes ou pequenas, sem que a escala assinale essa situação.

Ainda que não se possa dizer, rigorosamente, que a ordenação é uma medida, muitos a consideram como tal. A medida exige que a grandeza de um atributo seja expressa por uma unidade; entretanto, quando os objetos são ordenados segundo um atributo, nada se pode afirmar sobre a diferença em unidade de magnitude do atributo. É possível admitir, contudo, que a própria posição seja uma unidade, ainda que conceitualmente fraca, porque, muitas vezes, iguais diferenças de posição, como já foi assinalado, podem estar associadas .a diferenças desiguais na magnitude do atributo (JONES, 1971).

É possível, no caso da escala ordinal, usar qualquer simbologia, desde que expresse a posição dos indivíduos em suas relações uns com os outros. A informação que os números oferecem, conforme se viu, é bem limitada, refletindo, exclusivamente, a ideia de posição, sem permitir outras conclusões. Seria inteiramente sem sentido qualquer operação com os números de uma escala ordinal, pois nenhuma outra informação seria acrescentada à que já se possui (GUILFORD, 1965). As medidas educacionais, psicológicas, sociológicas etc. são expressas, geralmente, numa escala ordinal, salvo se alguns pressupostos forem admitidos e, nesse caso, teremos, então, escalas intervalares.

### 4.3. ESCALA INTERVALAR

Se além de distinguir diferenças entre as qualidades do atributo do objeto (medida ordinal), é igualmente possível estabelecer diferenças iguais entre essas propriedades, temos medidas numa escala de intervalo. Numa escala de intervalo, uma unidade é definida e o número atribuído à característica do objeto é igual ao número de unidades equivalentes à quantidade do atributo que o objeto possui.

A escala intervalar, além de refletir as propriedades da equivalência e da importância relativa, proporciona, também, uma medida o intervalo (distância) entre os valores da escala. É a primeira escala verdadeiramente quantitativa e tem como principal característica a existência de unidades ou intervalos constantes (ARMORE, 1967; SIEGEL, 1975). Às vezes, a referência a intervalos iguais pode levar à falsa suposição de que existe um número igual de pessoas ou objetos em cada ponto do *continuum*, mas, na verdade, o igual refere-se aos intervalos, independentemente do número de pessoas ou coisas em diferentes pontos da escala, conforme Nunnaly (1967). A diferença entre os escores 80 e 85, num teste de escolaridades, é igual à diferença entre 90 e 95, nesse mesmo teste, ainda que essas diferenças possam ter outras implicações.

A ocorrência de unidades iguais, na escala intervalar, possibilita estabelecer a diferença entre a posição dos indivíduos no atributo medido e comparar diferenças diversas umas com as outras. A medida do tempo (cronologia) e a de temperatura (escalas Centígrada e Fahrenheit) são exemplo de escalas de intervalo. O tempo decorrido entre 1930 e 1940 foi igual ao que decorreu entre 1960 e 1970. A diferença no atributo medido (tempo) é a mesma nos dois períodos, independentemente de sua localização na escala. Admitamos que as temperaturas médias, durante três dias, foram 18°C, 23°C e 33°C. O segundo dia foi 5°C mais quente do que o primeiro. As temperaturas nos dois primeiros dias foram mais semelhantes do que nos últimos dois dias, dois a primeira diferença (5° C) foi a metade da diferença da temperatura nos dois últimos dias (10° C). As diferenças, numa escala intervalar, possuem, portanto, significado.

O estabelecimento de razões entre medidas nesse tipo de escala é um procedimento que carece de sentido. Não se pode dizer que a temperatura de 40°C representa duas vezes mais calor do que a temperatura de 20°C ou que a variação de 30°C para

33°C significou um aumento de 10% de calor. Isso decorre de que o ponto zero na escala Centígrada foi fixado arbitrariamente e 0°C não significa ausência de calor.

A comparação entre a escala Kelvin, que possui um zero absoluto, indicativo da ausência de calor, e a escala Centígrada mostra, perfeitamente, porque não se pode estabelecer razões entre medidas numa escala intervalar. Admitamos os valores 0°, 50° e 100° na escala Centígrada. Os valores correspondentes, na escala Kelvin, são, respectivamente, 273°, 323° e 373°. Ambas as escalas usam as mesmas unidades para as suas diferenças, assim 50°C e 100°C correspondem a 323°K e 373°K, entretanto, um aumento de 50°C corresponde ao fator 2, enquanto que na escala Kelvin esse fator é 1,15. Isso decorre da posição do zero, que é diferente nas duas escalas (MINIUM, 1970).

As medidas na área educacional são, basicamente, ordinais, pois não indicam quantidades, mas ordens de posição dos indivíduos. As escalas ordinais não possuem intervalos iguais e nem zeros absolutos. A falta do zero absoluto na escala ordinal não é tão grave quanto a falta de intervalos iguais, pois mesmo não existindo o zero absoluto, é possível somar distâncias desde que os intervalos sejam iguais, conforme demonstra Kerlinger (1973), não sendo possível essa ocorrência sem que haja intervalos iguais, pelo menos teoricamente. Na área das ciências do homem – educação, sociologia, psicologia etc. – a maioria das escalas ordinais não pode ser considerada como possuindo intervalos iguais. Assim, se temos três medidas do mesmo traço, e essas medidas são substancialmente correlacionadas de modo linear, pode-se admitir que os intervalos sejam iguais (KERLINGER, 1973), numa visão pragmática do problema. Essa pressuposição é válida, pois quanto maior for a relação de linearidade, maior será, consequentemente, a possibilidade de se rem os intervalos iguais. Isso ocorre, geralmente, com os resultados de testes de escolaridade, testes e inteligência e escalas de atitudes. É possível que, ao considerar que as escalas ordinais tenham intervalos iguais, ocorram distorções e sejam introduzidos erros; contudo, se a construção do instrumento for cuidadosa e, especialmente, os resultados forem interpretados com as devidas cautelas, as consequências não têm amplas repercussões (KERLINGER, 1973).

A abordagem de Kerlinger (1973), nas suas linhas mais gerais, para transformar uma escala ordinal em escala intervalar,

assemelha-se à de Ghiselli (1964) e à de Guilford (1954), baseando-se essas abordagens nos seguintes argumentos: 1º) ainda que os procedimentos de construção dos testes não garantam uma escala de intervalo, pelo menos eles se aproximam do objetivo; 2º) tratar os escores de um teste como medidas de intervalo produz resultados úteis, possibilitando-nos, assim, admitir que temos uma escala de intervalo e agir como se tivéssemos esse tipo de escala, pois a análise dos resultados, segundo essa ótica, permite ter confiança nos pressupostos estabelecidos. As escalas de intervalo, conforme se verá, possuem inúmeras vantagens sobras escalas ordinais, daí a necessidade de expressar e interpretar os escores resultantes de testes como medidas de intervalo.

Os escores de um teste educacional, conforme discussão anterior são considerados como constituindo uma escala intervalar. Sem esse posicionamento, seria impossível, dada a natureza da sua escala (ordinal), estabelecer medidas de dispersão, como, por exemplo, a variância e o desvio padrão, que são indispensáveis para a definição de normas e a verificação do funcionamento efetivo do teste como instrumento de medida educacional.

Os sujeitos com mesmo escore num teste educacional são admitidos como possuindo igual capacidade. Um escore alto indica maior capacidade do que um escore baixo. Suponhamos, a título de exemplificação, os escores de três estudantes – 15, 20 e 30. É possível na escala intervalar, medir a diferença (distância) entre qualquer par de escores. Assim, no caso considerado, o segundo escore é cinco pontos maior do que o primeiro e o terceiro é maior 10 pontos em relação ao segundo escore. Além disso, nessa escala, a razão entre os intervalos tem significado. Consideremos os intervalos 5 e 10 entre o primeiro e o segundo escore, e entre este e o último escore. A razão entre os intervalos (2) indica que o terceiro escore (30) excede o segundo em duas vezes mais o que o segundo excede o primeiro. Observa-se, dessa forma, que a escala intervalar, além de especificar a equivalência, como na escala nominal, e a relação "maior do que", como na escala ordinal, específica, também, a razão de dois intervalos, quaisquer que sejam (SIEGEL, 1975).

Os escores dos testes de escolaridade, quando o número de itens é grande e a sua dificuldade bem distribuída, são tratados como uma escala de intervalo, conforme consideração anterior, a fim de possibilitar comparações inter e intraindivíduos.

Assim, tendo em vista a natureza da escala intervalar, não é possível afirmar que um escore 60 representa duas vezes mais conhecimento do que um escore 30, ou que as diferenças entre os escores 60 e 50 e entre os escores 15 e 5 significam a mesma diferença em rendimento escolar, ainda que essas diferenças sejam numericamente iguais. Malgrado os numerosos esforços na área da mensuração de variáveis educacionais, ainda não se conseguiu construir instrumentos que apresentem os resultados inequivocamente expressos numa escala de intervalo.

A grande limitação da escala intervalar centra-se no fato de não possuir um zero absoluto. O escore zero em um teste educacional não significa absoluta falta da capacidade medida por um. instrumento. O ponto zero, na escala de intervalo, é estabelecido de modo arbitrário.

### 4.4. ESCALA DE RAZÃO

A escala de razão, que é um tipo particular de escala de intervalo, possui todas as propriedades da escala intervalar mais o zero absoluto ou verdadeiro como origem. O zero absoluto significa ausência total do atributo mensurado. Graças ao zero absoluto, a escala de razão proporciona uma medida do intervalo de um certo valor em relação ao zero absoluto. Isso tem grande importância, pois a razão entre dois valores da escala é significativa assim como também é significativa a razão entre dois intervalos dessa escala.

Quando a escala não possui um zero absoluto, a soma das medidas não permite interpretações adequadas, pois, nesse caso, o valor numérico de uma medida representa uma distância a partir de uma origem arbitrária e inclui uma constante, geralmente de tamanho desconhecido, que representa a distância da origem arbitrária ao zero absoluto. Assim, quando são somadas duas dessas medidas, a soma inclui uma quantidade igual a duas vezes a constante desconhecida. Entretanto, ainda que a soma de duas medidas apresente dificuldades de interpretação, quando a escala não é de razão, a média de duas ou mais medidas pode ser interpretada do mesmo modo que as medidas individuais (JONES, 1971).

A maioria das medidas físicas – temperatura na escala Kelvin, comprimento, peso etc. – forma escalas de razão. As medidas nessa escala, além de refletirem diferenças na quantidade

do atributo (escala de intervalo), mostram, também, quantas vezes a quantidade do atributo é maior ou menor do que a quantidade do atributo de outro objeto.

O problema da mensuração de variáveis educacionais em escalas de razão ainda não foi satisfatoriamente resolvido. O zero num teste educacional não significa, necessariamente, ausência da capacidade mensurada, consequentemente, 75 itens respondidos corretamente não significam uma capacidade três vezes maior do que a capacidade representada por 25 itens também respondidos corretamente (GUILFORD, 1954). Entretanto, os escores dum teste podem ser considerados como formando uma escala de razão desde que o nosso interesse se limite, exclusivamente, ao número de itens respondidos corretamente. Isso não ocorre na maioria das vezes, porque se procura, na verdade, dar um significado ao escore, que é usado para indicar a posição do indivíduo numa escala de capacidade. E quando isso se verifica, as frequências da distribuição perdem suas propriedades de razão.

## 5. CARACTERÍSTICAS DAS VÁRIAS ESCALAS DE MEDIDAS – UM RESUMO

*Equivalência*

Os elementos são categorizados e as categorias representadas por números. Todos os elementos em uma categoria são equivalentes (iguais). O número de uma categoria é maior ou menor do que um outro número e nada diz sobre os atributos dos elementos, salvo que são iguais ou diferentes.

*Importância relativa*

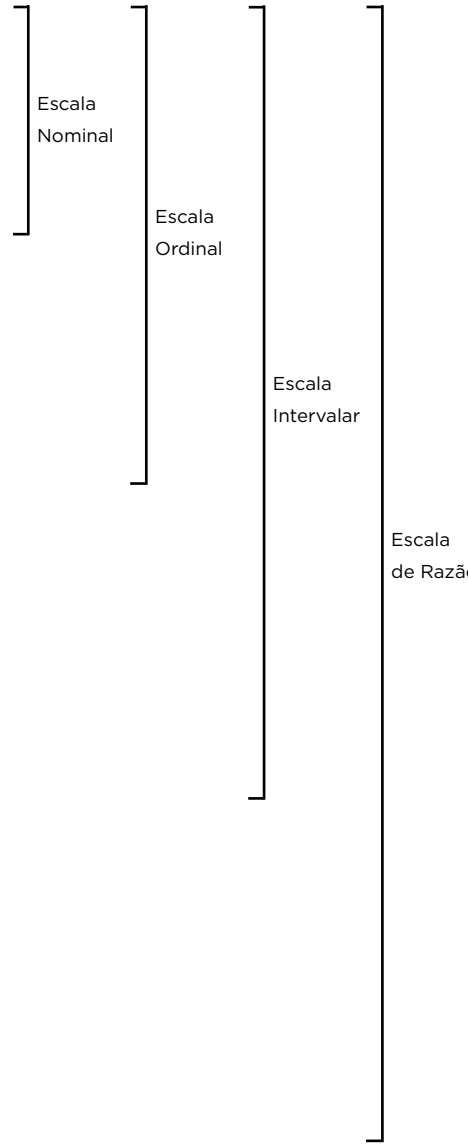
A grandeza relativa dos números atribuídos aos elementos reflete a quantidade do atributo possuída pelo elemento, indica a relação *maior do que* ou *menor do que*. Iguais diferenças entre os números não significam iguais diferenças nas quantidades dos atributos dos elementos.

*Proporciona uma medida do intervalo (distância) entre valores da escala*

Uma unidade de medida é fundamental para caracterizar essa escala; desse modo, os números, além de significarem ordenação, mostram que diferenças iguais entre os números refletem igual diferença na quantidade do atributo medido, ou seja, a razão entre os intervalos da escala é significativa. O ponto zero é arbitrário e não reflete ausência do atributo.

*Proporciona uma medida do intervalo (distância) de um dado valor a partir do ponto zero verdadeiro.*

Os números atribuídos aos elementos possuem todas as propriedades manifestas na escala intervalar, além do mais, a escala apresenta um zero absoluto, que indica ausência do atributo mensurado. As razões entre os números atribuídos aos elementos refletem razões entre as quantidades dos atributos medidos. A razão entre os *valores* da escala é significativa.



## 6. PROBLEMAS DAS MEDIDAS – OBSERVAÇÃO FINAL

As discussões dos vários níveis de medida partiram da colocação de Stevens (1946), em seu trabalho hoje considerado clássico. O assunto não é tranquilo, e muitos estatísticos e psicometristas não concordam com a sua tese central de que o nível de medida condicionaria as possíveis manipulações matemáticas e estatísticas dos números que refletem o atributo medido. Alguns radicalizaram o problema e desenvolveram uma fundamentação estatística associada aos conceitos de Stevens. A obra de Senders é um exemplo típico dessa posição e no seu livro (1958), que tanta controvérsia provocou à época de sua publicação, Senders declara, textualmente, que a organização do livro não é a usual, pois não apresenta as medidas estatísticas na ordem convencional, mas sim na que é determinada pela escala de medidas, afirmando, ainda, que um número crescente de estatísticas se toma disponível quando se procede da escala nominal para a de razão. Outros, ao contrário, e são em grande número, negam a validade da teoria de que uma escala de medida ditaria o tipo de procedimento a empregar, como mostram, por exemplo, Boneau (1961) e Anderson (1961) ao discutirem posições assumidas por Siegel (1975). Lord, em nota bem-humorada (1953), mostrou ser possível manipular estatisticamente quaisquer números, inclusive os números de camisas de jogadores de futebol (escala nominal). O problema, conforme se verifica em Lord (1953), consistiria na interpretação do significado dos resultados, pois mesmo operando com escalas nominais e ordinais os resultados possibilitam uma interpretação rigorosa (LORD, 1954).

Algumas vezes, há, realmente, necessidade de categorizar as escalas de medidas, a fim de compreender suas limitações, mas não se deve partir do pressuposto de que toda e qualquer medida pode ser enquadrada em um esquema rígido – escalas nominais, ordinais, de intervalo e de razão – e de que nada existe fora dessa categorização. Na verdade, existem também outras escalas, que são variações e combinações das quatro escalas básicas, mas que são de restrita importância e sem maior aplicação na área educacional. No que diz respeito às medidas educacionais (psicológicas, sociológicas etc.), uma posição rigidamente ortodoxa poderia conduzir a um caos, pois, conforme Burke (1953), as propriedades de um conjunto de números

como urna escala de medidas não deve ter nenhum efeito sobre a escolha das técnicas estatísticas para representar e interpretar os números, posição esta que se opõe à de Senders (1953). Vimos que os escores de um teste formam uma escala ordinal, mas podem ser considerados como uma escala de intervalo e, ainda, se limitarmos os objetivos, podem formar uma escala de razão. Não se pode dizer, portanto, que esses escores formem, estritamente, uma escala ordinal ou de intervalo. Há quem os classifique como uma escala de "quase intervalo" (GLASS E STANLEY, 1970). O problema nuclear residiria, portanto, no uso e na interpretação dessas medidas. É necessário usar de bom-senso para que as conclusões extraídas dos números não violentem os princípios fundamentais da lógica.

## 7. REFERÊNCIAS BIBLIOGRÁFICAS


ANDERSON, C. Scales and statistics: parametric and nonparametric. *Psychological Bulletin*, Washington, v. 58, n. 4, p. 305-316, Jul. 1961.

ARMORE, S. J. *Introduction to statistical analysis and inference for psychology and education*. New York: John Wiley and Sons, 1967.

BONEAU, C. A. A note on measurement scales and statistical tests. *American Psychologist*, v. 16, n. 1, p. 160-261, may 1961.

BROWN, F. G. *Principles of educational and psychological testing*. Illinois: The Dryden Press, 1970.

BURKE, C. J. Additive scales and statistics. *Psychological Review*, v. 60, n. 1 p. 73-75, Jan. 1953.

GHISELLI, E. E. *Theory of psychological measurement*. New York: McGraw-Hill Book, 1964.

GLASS, G. V.; STANLEY, J. C. *Statistical methods in education and psychology*. Eaglewood Cliffs, New Jersey: Prentice-Hall, 1970.

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill Book, 1954. _____. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1965.

JONES, L. V. The nature of measurement. In: THONDIKE, R. L. *Educational measurement*. Washington, DC.: American Council on Education, 1971.

KERLINGER, F. M. *Foundations of behavioral research*. 2th ed. New York: Holt, Rinehart and Winston, 1973.

LORD, F. M. On the statistical treatment of football numbers. *American Psychologist*, Washington, DC., v. 8, n. 12, p. 750-751, Dec. 1953.

_____. Further comments on "football numbers". *American Psychologist*, v. 9, n. 6, p. 264-65, Jun. 1954.

LORGE I. The fundamental nature of measurement, In: LINDQUIST, E. F. *Educational Measurement*. Washington, D C. American Council on Education, 1951.

MAGNUSSON, D. *Test theory*. Reading, Mass: Addison-Wesley, 1967.

MINIUM, E. W. *Statistical reasoning in psychology and education*. New York: John Wiley and Sons, 1970.

NUNNALLY, J. C. *Psychometric Theory*. New York: McGraw-Hill Book, 1967.

_____. *Test and measurement*: assessment and prediction. New York: McGraw-Hill Book, 1959.

SENDERS, V. L. comment on burke's additive scales and statistics. *Psychological Review*, v. 60, p. 423-424, 1953.

_____. *Measurement and statistics*. New York: Oxford University Press, 1958.

SIEGEL, S. *Estatística não paramétrica para as Ciências do Comportamento*. São Paulo: McGraw-Hill, 1975.

STEVENS, S. S. On the theory of scales of measurement. *Science*, n. 103, p. 677-680, 1946.

# VALIDADE DE CONSTRUTO EM TESTES EDUCACIONAIS[1]

## INTRODUÇÃO

O problema da validade de construto[2] é de grande relevância na área educacional, tendo em vista o fato de que a avaliação se vale, frequentemente, de construtos, que, após sua definição operacional, são medidos por intermédio de testes. A *validade de conteúdo* e a *validade de critério* (concorrente e preditiva), atributos exigidos dos bons testes de escolaridade, não se preocupam, entretanto, com a compreensão dos construtos que os testes medem; consequentemente, impõe-se, conforme acentua Brown (1970), uma nova abordagem para análise dos instrumentos de medida da aprendizagem escolar.

A "validade" de muitos instrumentos de medida é, às vezes, inferida, mas tal procedimento é adotado à revelia da metodologia científica, não sendo, pois, um comportamento justificável. São necessárias provas insofismáveis de que um teste, construído para determinado fim, é efetiva mente válido. Desse modo, se um teste visa a medir processos mentais complexos, como ocorre inúmeras vezes na área educacional, inclusive nos exames de acesso às universidades, é indispensável que existam provas irrefutáveis de que os instrumentos utilizados medem, efetivamente, o construto hipotetizado.

**1** Artigo publicado em *Educação e Seleção*, n. 8, p. 35-44, jul./dez. 1983

**2** Constructos são traços, aptidões ou características supostamente existentes e abstraídos de uma variedade de comportamentos que tenham significado educacional (ou psicológico). Assim, fluência verbal, rendimento escolar, aptidão mecânica, inteligência, motivação, agressividade, entre outros, são constructos..

A validade de construto possibilita determinar qual a característica educacional que explica a variância do teste ou, então, qual o significado do teste (KERLINGER, 1973). Um avaliador, desse modo, poderá fazer perguntas do tipo: – o teste mede, efetivamente, a capacidade de expressão escrita? O teste é um reflexo do *status* socioeconômico dos estudantes avaliados? A indagação, no caso, pretende esclarecer qual a proporção da variância total do teste que decorre desses construtos: – expressão escrita e *status* socioeconômico. Ou, ainda, procura explicar as diferenças individuais nos escores desse instrumento de medida. O interesse, na validação de construtos, centraliza-se na característica ou traço que está sendo medido, mais do que no próprio teste.

A verificação da validade de construto, sua lógica e metodologia foram amplamente estudadas por Cronbach e Meehl (1955), que produziram documento fundamental para a compreensão desse tipo de validade, que é de particular importância sempre que um instrumento deva ser interpretado como proporcionando medidas de um atributo ou qualidade que, presumivelmente, as pessoas possuem. Outro documento básico para o estudo do problema é o ensaio de Campbell e Fiske (1959) sobre validade convergente e discriminante. Ambos os trabalhos são amplamente utilizados na fundamentação do presente estudo.

## O CONCEITO DE VALIDADE DE CONSTRUTO

A validade de construto, ao contrário da validade empírica, não é expressa em termos de um coeficiente quantitativo, conforme ocorre no caso da validade preditiva e da validade concorrente. O conceito de validade de construto, por sua vez, é extremamente útil para explicar a natureza dos instrumentos que medem traços para os quais não se possuem critérios externos. Assim sendo, é necessário partir de uma variável logicamente definida. A variável, como um construto lógico, é inserida num sistema de conceitos, cujas relações são explicadas por uma teoria e a partir da qual certas consequências práticas, sob determinadas condições, podem ser extraídas e testadas (MAGNUSSON, 1967). Se o resultado é o que se esperava em uma série de testes, o instrumento é considerado como possuindo validade de construto para a variável testada. Assim, a constatação

da validade de construto resulta do acúmulo, por diferentes meios, de várias provas, que precisam ser analisadas em todos os seus detalhes, a fim de constatar, entre outros aspectos, quais as variáveis com as quais os escores do teste se correlacionam, quais os tipos de itens que integram o teste, o grau de estabilidade dos escores sob condições as mais variadas e o grau de homogeneidade do teste, com vistas a ter elementos que possam esclarecer o significado do instrumento.

A compreensão de um instrumento que mede determinado construto ganha nova dimensão quando se conhece o grau de relacionamento com outros construtos. Exemplificando, os escores de um teste especialmente construído para medir "rendimento escolar" adquire novo sentido quando se estabelece o seu grau de associação com outros construtos, como "inteligência" e "criatividade". A validade de construto possibilita determinar se o teste é a melhor medida de algo diferente do que foi pretendido medir ou, então, se fornece elementos que não possuem efetiva consequência no processo de avaliação.

## AS TEORIAS EDUCACIONAIS E A LÓGICA DA VALIDADE DE CONSTRUTO

É comum em educação (e em psicologia) o desenvolvimento de sistemas unificados de princípios, definições, postulados e observações, para explicar o relacionamento entre variáveis (SAX, 1980), ou seja, o desenvolvimento de teorias educacionais, a partir das quais são construídos instrumentos para a mensuração de um determinado traço ou característica. Somente por intermédio da constatação da validade de construto desses instrumentos é que se pode confirmar o significado dessas características ou traços apresentados pela teoria. E para alcançar esse objetivo, necessário se faz a aplicação dos procedimentos clássicos do método dedutivo: teoria; dedução, hipótese, experimentação e, finalmente, dados que confirmem ou neguem a hipótese, ou seja, o construto.

A lógica da validade de construto, assim como o seu processo, são, essencialmente, os do método científico. Parte-se de uma teoria sobre a natureza do construto e fazem-se predições sobre as relações entre os escores do teste e outras variáveis. A seguir,

essas predições são verificadas empiricamente e, dependendo dos resultados, a teoria é aceita ou revista. O trabalho contínuo de fazer predições, testar hipóteses, através da experimentação, e rever a teoria são atividades que contribuem para aumentar a precisão da definição do construto.

## A REDE NOMOLÓGICA, SEGUNDO CRONBACH E MEEHL

As teorias que procuram explicar um determinado fenômeno compreendem um conjunto interrelacionado de conceitos, proposições e leis. A esse sistema interligado Cronbach e Meehl (1955) deram o nome de rede nomológica[3]. As leis estabelecem relações entre diferentes características, entre características e construtos, ou entre um construto e outro, conforme aqueles autores, os quais ressaltam ainda que as leis e os conceitos devem estar ligados a comportamentos observáveis. Às vezes, entretanto, isso não ocorre. Uma determinada lei ou conceito não se relaciona diretamente com as características, mas com outras leis e conceitos, que, por sua vez, são diretamente ligados a características observáveis. É condição necessária, na pesquisa da validade de construto, que a definição de conceitos ou formulação de leis estejam apoiados, direta ou indiretamente, em dados observáveis.

**3** Nomológico - relativo à nomologia, estudo das leis que presidem aos fenômenos naturais.

Cronbach e Meehl (1955) assinalam, ainda, que o significado de um conceito somente fica perfeitamente esclarecido quando se elabora uma rede de relações que mostre, claramente, que essas relações são específicas e definitivas. Dessa forma, na validação de um construto, há necessidade de um grande número de diferentes operações, inclusive de ordem qualitativa, para a mensuração de um conceito e a fim de mostrar que esse mesmo conceito está ligado a outro conceito por intermédio de uma rede nomológica.

A partir desse posicionamento, é preciso considerar as seguintes situações:

1. o processo de inferência do significado de um construto exige que os dados sejam observáveis;
2. o processo de inferência do significado de um construto, a partir de dados observáveis, deve ser explicitamente

especificado, para que se possa verificar a precisão de uma inferência;

3. diferentes usuários de um construto devem servir-se, essencialmente, de uma mesma rede nomológica, para que a concordância entre os pesquisadores seja possível, pois, frequentemente, o nome de um mesmo construto é usado com referência a diferentes construtos (ex: inteligência) ou diferentes nomes são empregados com relação ao mesmo construto (ex: pensamento divergente e criatividade), conforme assinalou Brown (1970).

## VALIDAÇÃO DE TESTES E VALIDAÇÃO DE TEORIAS

A validação de construtos não se limita apenas a validar um teste, o seu alcance é bem mais amplo, centrando-se o seu objetivo na validação da teoria em que se apoiou a construção do instrumento (KERLINGER, 1973); desse modo, o trabalho de validação de um construto é urna pesquisa científica empírica, porque, definidos os construtos que seriam responsáveis pelo desempenho no teste o avaliador passa a formular hipóteses sobre a teoria dos construtos e, a seguir, testa empiricamente essas hipóteses. A testagem de hipóteses sobre construtos exige a verificação da *convergência* e da *discriminatividade*. A *convergência* mostra que as evidências coletadas de diferentes fontes e de diferentes modos indicam um significado igual ou semelhante para o construto. A *discriminatividade* refere-se à possibilidade de diferenciar, empiricamente, um construto de outros construtos semelhantes, assim como a de constatar o que não está correlacionado aos construtos (KERLINGER, 1973). As ideias de *convergência* e de *discriminatividade* serão retomadas mais adiante, quando for discutida a metodologia de Campbell e Fiske (1959) relativa à matriz do multitraço-multimétodo. A validação de construtos ultrapassa, assim, os limites de uma validação empírica, pois, além de constatar a correlação com um critério ou em que medida o instrumento separa indivíduos que possuem alto ou baixo grau de uma determinada característica, explica, também, o porquê dessas ocorrências.

## O PROCESSO DE VALIDAÇÃO E TESTAGEM DE HIPÓTESES

O processo de validação exige, essencialmente, o estabelecimento de hipóteses, a partir de leis e construtos definidos pela rede nomológica, e a coleta de dados para testar essas hipóteses. Ao discutir os dados sobre a validade de um construto é necessário que sejam claramente. especificados os seguintes aspectos (BROWN, 1970):

1. a interpretação proposta, ou seja, qual o construto que se tem em mente, como esse construto é definido e como a hipótese testada foi estabelecida a partir de uma teoria importante;
2. a comprovação adequada da interpretação, oferecendo elementos que apoiaram ou rejeitaram as hipóteses;
3. a argumentação sobre a concretização dos objetivos propostos (apresentar detalhes sobre os procedimentos experimentais e a linha de raciocínio que permitiu inferências sobre o significado do construto).

O que significam, efetivamente, os resultados da comprovação das hipóteses levantadas? Se as predições forem confirmadas empiricamente, pode-se acreditar que o teste meça o construto e ter maior confiança no conceito adotado. Um construto nunca pode ser comprovado como correto em termos absolutos (CRONBACH; MEEHL, 1955), mas somente adotado como a melhor definição de trabalho. Se, ao contrário, os resultados forem negativos e os dados não confirmarem as predições, três interpretações são possíveis: 1º – o teste não mede o construto; 2º – o referencial teórico não é correto, possibilitando inferências errôneas; e 3º – o planejamento experimental não possibilita a testagem de hipóteses.

O fracasso na confirmação de uma predição indica a necessidade de uma revisão na rede teórica ou no procedimento experimental (BROWN, 1970). As interpretações ambíguas de resultados negativos constituem uma das desvantagens da validade de construto.

## A METODOLOGIA DA PESQUISA DA VALIDADE DE CONSTRUTO

A validade de construto pode ser pesquisada por diferentes métodos, inclusive os que são empregados na validade de conteúdo e de critério. Ao utilizar diferentes métodos é importante que se estude (MAGNUSSON, 1967), entre outros aspectos:

a) as diferenças entre os grupos em função do que a teoria estabelece relativamente à variável pesquisada;
b) como os resultados dos testes são influenciados por mudanças nos indivíduos ou no meio, as quais, segundo a teoria, devem, respectivamente, influenciar ou deixar de influenciar as várias posições dos indivíduos no *continuum*;
c) as correlações entre diferentes testes que, supostamente, medem a mesma variável. É necessário cautela a fim de que as correlações entre as medidas não se elevem em virtude de similaridades nos métodos utilizados (CRONBACH; MEEHL, 1955). Isso pode acontecer (MAGNUSSON, 1967) se as respostas dos testes exigirem alguma aptidão especial além da que está sendo considerada. Uma possível concordância entre as medidas poderia, simplesmente, provocar um aumento da correlação do efeito de diferenças individuais relativas a essa aptidão especial;
d) as correlações entre itens isolados ou diferentes partes do teste, a fim de verificar se possuem alta intercorrelação e o teste possa ser considerado como medindo uma variável unitária.

A validade em geral e, especialmente, a validade de construto são as estimadas pela concordância de medidas obtidas por métodos tão diferentes quanto possível. As dissimilaridades dos métodos, no estudo da validade de construto, são importantes para que as intercorrelações obtidas possam ser interpretadas como expressando realmente esse tipo de validade.

## A VARIÂNCIA NO PROCESSO DE VALIDAÇÃO: MÉTODOS E INDIVÍDUOS

Deve-se assinalar que a variação entre os indivíduos é expressa por escores que foram obtidos por um determinado método pré-definido, podendo esses escores ser afetados por diferenças, ainda que irrelevantes, nas reações dos indivíduos às características do método ou, então, por diferenças nas posições ao longo do *continuum* que se pretende que o teste meça, ou, ainda, por ambas as situações (CRONBACH; MEEHL, 1955; MAGNUSSOM, 1967). Assim, uma parte da variância total da distribuição dos escores pode ser atribuída a aspectos especificamente característicos do método empregado na mensuração, enquanto outra parte da variância pode ser considerada como resultado de diferenças realmente verdadeiras entre os indivíduos no que diz respeito ao traço mensurado. Esta variância expressa a *variância verdadeira* é a que se deseja determinar com o máximo de precisão na pesquisa da validade. Desse modo, pode-se decompor a variância sistemática em:

a) variância devida às propriedades do método empregado, e
b) variância decorrente de características relevantes dos indivíduos testados.

A variância resultante de propriedades do método empregado, conforme acentua Magnusson (1967), gera uma espécie de efeito de halo[4] metodológico. Desse modo, quando as medidas de um certo número de variáveis são baseadas em um único método, os coeficientes da matriz de intercorrelação são grandes, em geral. A medida de diferentes traços é, portanto, afetada pelas propriedades do método empregado, as quais contribuirão para que resulte uma certa quantidade de variância comum.

É preciso atentar para o fato de que, quando dois métodos diferentes mas com propriedades semelhantes são independentemente empregados na medida de certo traço, uma parte da variância das medidas baseadas em um método pode repetir-se, sistematicamente, na medida baseada no outro método (MAGNUSSON, 1967). A variância comum, resultante de semelhanças entre os métodos empregados, redundará numa superestimação da validade de construto, quando esta é verificada pela intercorrelação de escores obtidos a partir de diferentes métodos.

**4** O *efeito de halo* é um efeito sistemático do avaliador que deve ser levado em conta quando traços humanos estão sendo avaliados. Uma atitude positiva ou negativa do avaliador em relação ao avaliado afeta, geralmente, na direção da atitude, as medidas de cada traço sujeito à avaliação. O efeito poderá provocar um decréscimo nas diferenças individuais e um aumento na correlação entre as medias de diferentes traços.

## MÉTODOS USADOS NA VALIDAÇÃO DE CONSTRUTOS

Os métodos usados no estudo da validade de construto podem ser classificados, de acordo com Brown (1970), em cinco categorias: métodos intratestes, métodos entretestes, estudos relacionados a critérios, estudos experimentais e estudos de generalizabilidade[5].

**5** Generalizabilidade – propriedade que têm as coisas de se tornarem generalizáveis.

*Métodos intratestes* – esses métodos usam técnicas que permitem o estudo da estrutura interna do teste – seu conteúdo, as interrelações entre os itens e os subtestes e os processos relacionados com as respostas aos itens. Essas técnicas não consideram variáveis externas, porque se preocupam, antes de mais nada, com a estrutura interna do teste. Assim, não se pode usar essa metodologia para obter amplas informações sobre a validade de construto do teste, quando muito o seu emprego possibilitaria algum conhecimento sobre a natureza do construto, mas não o relacionamento do construto com outras variáveis.

A determinação da *validade de conteúdo* fornece informações sobre a validade de construto e é um tipo de estudo que pode ser incluído na categoria dos métodos intratestes. As especificações do conteúdo e do domínio comportamental "amostrado" no teste, condição essencial no estudo da validade de conteúdo, também servem para definir a natureza do construto que o teste mede.

A análise da *homogeneidade do teste*, método que também pode ser incluído na categoria dos intratestes, por intermédio de medidas de consistência interna (coeficientes de Kuder Richardson), pela análise fatorial dos itens, entre outros estudos de homogeneidade, auxiliam na definição do construto, especialmente ao indicar se o teste mede um único traço ou se, ao contrário, mede diversos traços.

Ao analisar um teste, o interesse nem sempre se limita ao conhecimento do conteúdo dos itens, aprofunda-se e procura conhecer, também, o processo usado pelos examinandos nas suas respostas aos itens. Assim sendo, qualquer processo de análise que identifique capacidade e habilidades pode, em princípio, esclarecer o significado do construto que o teste mede, ao indicar as variáveis que estão sendo medidas pelos itens do instrumento. É preciso, entretanto, usar de cautela quando for empregado o processo de análise, pois, indiscutivelmente, diferentes pessoas poderão utilizar processos diversos, mas igualmente válidos, de

resposta a um item, criando-se, desse modo, uma situação complexa que pode levar a falsas interpretações.

*Métodos entretestes* – os métodos incluídos nessa categoria consideram, simultaneamente, vários testes, mas não levam em consideração variáveis extratestes. Os métodos entretestes permitem indicar, geralmente, os aspectos comuns a vários testes, mas não possibilitam a realização de inferências diretas sobre a relação entre escores do teste e variáveis externas.

O método mais simples dessa categoria consiste em correlacionar um teste novo a um outro teste já amplamente estudado e conhecido nas suas diversas dimensões. É a chamada *validade congruente*. Se existe essa alta correlação entre os dois testes pode-se dizer que ambos medem o mesmo construto. Essa abordagem apresenta um aspecto que merece reflexão, pois a menos que os dois testes sejam altamente correlacionados (BROWN, 1970), isto é, as correlações e as fidedignidades sejam da mesma magnitude, os fatores que influenciam no abaixamento da correlação podem ser importantes para determinar a relação entre o teste e a variável externa e, desse modo, invalidar as inferências realizadas com base na intercorrelação dos testes.

Uma outra abordagem, ainda nessa categoria, consiste em promover a *análise fatorial* em um grupo de testes. A análise fatorial é, no momento, o método mais promissor para estimativa da validade de construto (KERLINGER, 1970), pois objetiva reduzir um grande número de medidas a um número reduzido de fatores a fim de estabelecer quais os que medem as mesmas coisas e em. que medida está, realmente, ocorrendo a mensuração que era esperada. A análise fatorial, ainda que seja um procedimento complexo, exigindo inclusive o emprego de computador, é um caminho fecundo na pesquisa de construtos, pois mostrará quais os testes que compartilham da variância comum e, assim, medem o mesmo construto pesquisa das cargas do mesmo fator (BROWN, 1970), no conteúdo comum dos testes, possibilita inferir sobre a natureza do construto e até mesmo pode levar à identificação do fator. A análise fatorial permitirá verificar em que medida cada teste está saturado pela variância comum e em que medida os seus escores dependem da variância específica. A proporção da variância total dos escores do teste, que é variância comum, é um índice da validade de construto. O presente método visa a estabele-

cer a *validade fatorial*, denominação esta frequentemente usada como sinônimo de validade de construto.

Outra abordagem possível nessa categoria refere-se às concepções de Campbell e Fiske (1959) para estabelecimento da *validade convergente* e da *validade discriminante*. A validação, segundo estes autores, processar-se-ia por meio de métodos que visariam a estabelecer se duas técnicas de medida (ou testes) diferentes estariam medindo, efetivamente, o mesmo construto; daí serem esses métodos chamados de convergentes. A *validade congruente* e a *validade fatorial* são exemplos típicos do uso de métodos convergentes.

Um dos problemas centrais da análise da validade de um construto consiste em que os testes, além de apresentarem uma alta correlação com outros testes que medem o mesmo construto, devem, também, demonstrar que não apresentam correlação com testes que medem claramente construtos diferentes. Sabe, por exemplo, que há uma associação entre inteligência e criatividade; dessa forma, segundo o ponto de vista de Campbel e Fiske (1955), somente podemos aceitar um teste de criatividade se o mesmo não apresentar qualquer correlação com os resultados de testes de inteligência, quando se terá certeza de que aquele construto está sendo medido e não se confunde com este outro. Esse é, sem dúvida, o ponto discriminante na validação de construto, pois, conforme acentuou Brown (1970), um teste somente é válido para medir um determinado construto quando, sem sombra de dúvida, é independente de testes que medem outros construtos perfeitamente definidos.

*Estudos relacionados a critérios* – A natureza e o tipo de critérios que os escores de um teste predizem são indicativos do construto que o teste mede; isso posto, é perfeitamente possível obter importantes informações para o estabelecimento da validade de construto a partir dos dados de estudos sobre validade de critério.

Uma das maneiras de coletar evidências sobre um construto é a partir de escores de um teste que seja capaz de separar grupos naturalmente existentes de grupos organizados experimentalmente. Os escores para esse fim devem estabelecer a diferenciação entre grupos. Suponhamos, a título de exemplificação, que foi construído um teste para identificar criatividade literária. Os itens desse teste devem ter sido elaborados de forma a identificar dois grupos bem distintos: os que são capazes de

produzir textos de valor artístico nos vários setores da literatura e aqueles que apenas possuem o domínio da língua, como qualquer pessoa comum. O instrumento assim construído teria validade de critério (concorrente), o que representa importante informação para a caracterização do construto.

Uma outra abordagem, ainda segundo essa perspectiva, seria a formação de dois grupos distintos com base nos escores do teste (quartil superior e inferior, por exemplo) e o estabelecimento das características de cada um desses grupos, o que permitiria estabelecer uma definição tão completa quanto possível, sobre a natureza do construto.

Destaca-se, nessa categoria, o método baseado no emprego de *coeficientes de validade*, o qual consiste em verificar o êxito de um instrumento na predição de um determinado comportamento. Um teste de aptidão escolar deve ser um bom preditor do desempenho acadêmico, pois há uma clara associação entre essas duas variáveis. Ora, na medida em que isso ocorre, o teste critério estaria medindo se o construto é realmente aptidão escolar.

*Estudos experimentais* – Outros métodos usados para determinação da validade de construto exigem a manipulação de algumas variáveis e a observação dos efeitos consequentes nos escores do teste. Tomem a. o exemplo da ansiedade, conforme a colocação de Brown (1970). Os estudantes, em época de exames, costumam demonstrar um aumento na sua ansiedade. A ansiedade durante a realização de um teste refletiria medo de fracasso no exame e o comprometimento do autoconceito da pessoa. A partir dessa definição de ansiedade, pode-se estabelecer a hipótese de que o desempenho no exame é inversamente correlacionado à ansiedade durante a realização de um teste, se o teste for de grande importância para a pessoa. Igualmente, pode-se hipotetizar que ansiedade e exame não se correlacionam, se o teste for estruturado de forma a não constituir um11 ameaça ao indivíduo. Se, nesse contexto, os escores num teste de ansiedade não apresentarem relação de predição com os escores de um exame, ter-se-ia, então, evidência de que o teste mede realmente o construto ansiedade.

Um conceito de grande importância em psicometria – o de fidedignidade – pode ser usado no estudo da validade de construto (CRONBACH; MEEHL, 1955). Se o construto estabelece que o traço a ser medido é grandemente estável ao longo do

tempo e resiste a mudança. O coeficiente de estabilidade do teste será um indicador do construto que foi hipotetizado. Se o construto estabelece que os escores, em certas circunstâncias, aumentam com a idade, na medida em que isso é constatado, tem-se uma prova da validade de construto de instrumento.

*Estudos de generalizabilidade: método do multitraço-multimétodo* – Os estudos de generalizabilidade são aqueles em que o teste, cujo construto se deseja validar, é analisado sob diferentes condições, como, por exemplo, a aplicação do instrumento a diferentes amostras da população, a utilização de métodos diversos na sua aplicação etc. A abordagem mais frequentemente utilizada para estudos desse tipo é o da matriz do *multitraço-multimétodo* conforme a proposta de Campbel e Fiske (1959).Esses autores estabeleceram algumas condições fundamentais para o sucesso de um processo de validação. Essas condições não se restringem apenas à validade de construto, incluindo, também, a verificação da validade preditiva e concorrente.

Um teste é uma unidade traço-método (CAMPBEL; FISKE, 1959), ou seja, um teste mede determinado traço usando um único método. O interesse, portanto, no processo de validação, centra-se no conhecimento das contribuições relativas do traço e do método para o escore do teste, o que obriga a estudar mais de um traço e mais de um método. Isso significa que se deseja, na verdade, estabelecer a *validade convergente* e a v*alidade discriminante*.

A *validade convergente* será determinada por intermédio da correlação entre os *mesmos traços* medidos por *diferentes métodos* esperando-se que essa correlação seja significantemente alta. A *validade discriminante,* por sua vez, será estabelecida comprovando-se que *diferentes traços*, mesmo quando medidos pelo *mesmo método*, não possuem uma alta correlação.

A Tabela 1.0 apresenta uma matriz relativa ao emprego do *multitraço-multimétodo*, em que temos três traços hipotéticos (A, B, C) medidos por três métodos diferentes (1, 2, 3), que geram nove variáveis separadas. Ou mais claramente, a título de exemplificação, os traços seriam: – compreensão de textos, raciocínio abstrato e capacidade de identificar elementos secundários numa informação, traços esses que seriam medidos por um teste de papel-e-lápis, um teste individual e uma escala de classificação. O método pode ser usado para o estudo de n traços, usando *m* métodos, não havendo necessidade de n = *m*.

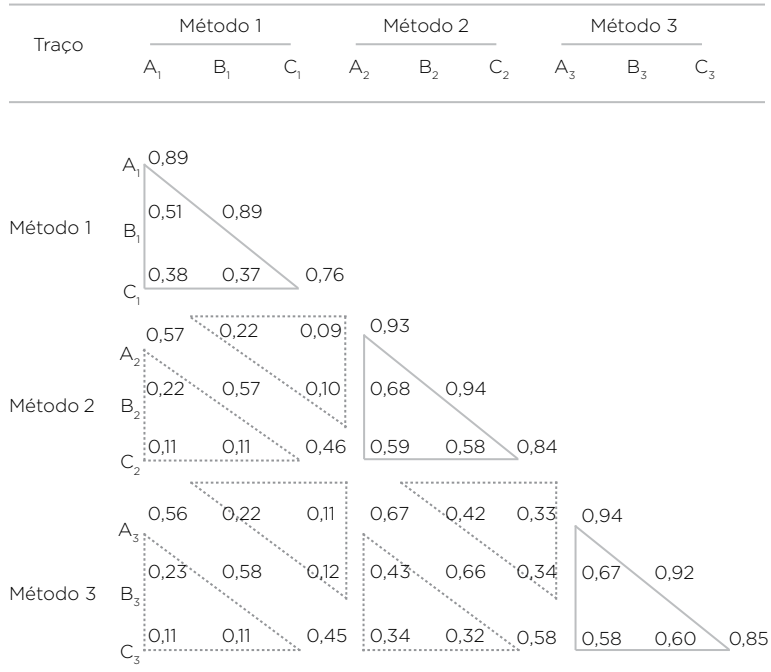
O número de traços é igual ao número de métodos, no exemplo citado, apenas por conveniência na discussão da metodologia.

Os coeficientes de correlação apresentados na Tabela 1.0 representam o grau de associação entre *três variáveis* (A, B, C), medidas por *três métodos diferentes* (1, 2, 3). Os escores para cada uma das variáveis são correlacionados com os escores de cada uma das outras variáveis, independentemente do método pelo qual os escores foram obtidos. Os valores na diagonal da matriz completa – 089; 0,89; 0,76;.....0,94; 0,92; 0,85 – representam as *fidedignidades das medidas*, valores esses que representam os resultados da medida do *mesmo traço* pelo *mesmo método*: portanto, são os valores *monotraço-monométodo*.

Os triângulos representados por linhas cheias, ao longo da diagonal da matriz completa, contêm coeficientes que dão a relação entre as medidas de *diferentes variáveis* pelo *mesmo método*, são, pois, valores *heterotraço-monométodo*. Considerando-se que o mesmo método foi usado para a medida de diferentes variáveis, as propriedades do método dão origem a uma variância comum para as diferentes variáveis, na medida em que as propriedades do método concorrem para a variância sistemática quando variáveis individuais são mensuradas.

Os triângulos em linhas interrompidas contêm coeficientes de correlação entre medidas de *diferentes variáveis* obtidas por *métodos diferentes*; dessa forma, são valores *heterotraço heterométodo*. As diagonais entre esses triângulos apresentam os coeficientes de correlação entre as medidas da *mesma variável* por *diferentes métodos* e são *coeficientes de validade* (validade convergente). Esses valores em diagonal devem ser substanciais, porque refletem a correlação entre as mesmas variáveis medidas diferentemente. A variância comum que resulta, em virtude das semelhanças nos métodos, afetará o tamanho desses coeficientes na medida em que os métodos têm propriedades iguais e lhes são oferecidas oportunidades de afetar as medidas de maneira sistemática.

**TABELA 1.0. Matriz multitraço-multimétodo, segundo Campbel e Fiske (1959)**




## MÉTODO DO MULTITRAÇO-MULTIMÉTODO: CONDIÇÕES PARA APLICAÇÃO

Campbel e Fiske (1959) estabeleceram algumas condições para, usando o método do multitraço-multimétodo, desenvolver o processo de validação:

**1º.** os coeficientes de correlação entre medidas da *mesma variável* com *diferentes métodos* (coeficiente de validade convergente) devem ser significativamente maiores do que zero.(Este critério é, normalmente, considerado suficiente para caracterizar a validade);

**2º.** as medidas de uma variável devem apresentar uma correlação mais estreita com medidas do *mesmo tipo*, e que foram obtidas por um *outro método*, do que com medidas de *outro tipo* que foram estabelecidas pelo *mesmo método*. Os coeficientes de validade para uma certa variável devem, assim, ser maiores do que os coeficientes

para a mesma variável nos triângulos delimitados por linhas contínuas;

3º. o coeficiente de validade para uma determinada variável deve ser maior do que a correlação entre as medidas dessa variável e as medidas de todas as outras variáveis, obtidas por qualquer outro método. Um coeficiente de validade, desse modo, será maior do que os correspondentes coeficientes na mesma linha e coluna no interior do triângulo representado por linhas interrompidas;

4º. se os mesmos métodos ou métodos diferentes forem usados, as magnitudes dos coeficientes para as correlações entre diferentes variáveis devem ter a mesma configuração.

Se a primeira condição for satisfeita, os métodos possuem *validade convergente*: concordância significativa entre medidas do mesmo tipo com diferentes métodos. A concretização dessa condição não é suficiente para satisfazer o processo de validação. É necessário que a segunda e a terceira condição também ocorram. Se ambas ocorrerem, dir-se-á que as medidas têm *validade discriminante*.

É preciso, no caso das condições estabelecidas por Campbell e Fiske (1959), considerar a fidedignidade dos métodos, pois se ocorrer a falta desse atributo, a validade discriminante será afetada. A fidedignidade, na presente abordagem, refere-se à concordância entre duas medidas do *mesmo traço* usando o *mesmo método* enquanto que a *validade* é definida em termos da concordância entre duas medidas do *mesmo traço* usando *diferentes métodos*. Assim sendo, cumpre ressaltar que a diferença fundamental entre fidedignidade e validade está na similaridade dos métodos de medida. A quarta condição apresentada por Campbell e Fiske (I959) é irrealista (MAGNUSSON, 1967) e impossível de ser obedecida rigorosamente. Se fosse seguida, a validade discriminante dificilmente seria estabelecida, em face da dificuldade de julgar o efeito da falta de fidedignidade em matrizes complexas, com inúmeras variáveis, como exigem os estudos de validação de construtos.

## REFERÊNCIAS BIBLIOGRÁFICAS


BROWN, F.G. *Principles of Educational and psychological testing*. Hinsdale, lllinois: The Dryden Press, 1970.

CAMPBELL, D. T.; FlSKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, n. 59, 1959.

CRONBACH, L. J.; MEEHL, P. E. Construct validity and psychological tests. *Psychological Bulletin*, n. 52, 1955.

KERLINGER, F. M. *Foundations of Behavioral Research*. 2th ed. New York: Holt, Rinehart and Winston, 1973.

MAGNUSSON, D. *Test theory*. Reading, Mass: Addison-Wesley, 1967.

SAX, G. *Principles of educational and psychological measurement and evaluation*. 2th ed. California: Wadsworth Publishing, 1980.

# APLICAÇÃO DE CRITÉRIOS DE CORREÇÃO EM PROVAS DE REDAÇÃO[1]

## INTRODUÇÃO

A partir de 1965, a seleção de candidatos para algumas universidades e escolas superiores brasileiras começou a ser feita por meio de provas objetivas, com o emprego de itens de múltipla escolha. Aos poucos, o uso desse tipo de instrumento de medida educacional se difundiu e, na década de 70, o que representava uma experiência limitada a determinados centros educacionais passou a traduzir o comportamento geral das instituições de ensino superior, com apoio em normas oriundas do Ministério da Educação e Cultura. Isso significou, consequentemente, o abandono da redação, instrumento de seleção tradicionalmente empregado nos vestibulares, desde a criação desses concursos em 1911.

A reação ao uso exclusivo de provas objetivas não se fez esperar. A Universidade de São Paulo, em 1976, reintroduziu, em seus exames de ingresso, provas dissertativas, que, a partir de janeiro de 1978, também passaram a ser adotadas pelas demais instituições de ensino superior do país, como decorrência da nova orientação do Ministério da Educação e Cultura, que tomou obrigatória a prova de redação em todos os vestibulares.

**1** Artigo publicado em *Cadernos de Pesquisa*, n. 26, p. 29-34, set. 1978.

A redação, na atual sistemática dos vestibulares, que envolvem, muitas vezes, mais de 100.000 estudantes, como nos casos de São Paulo e do Rio de Janeiro, criou, naturalmente, numerosos problemas, entre os quais sobressai o da aplicação de um critério de correção, objetivo do presente estudo.

## ELABORAÇÃO DO CRITÉRIO DIRETRIZES GERAIS

A Fundação Carlos Chagas, no primeiro semestre de 1975, iniciou estudos e procurou equacionar, com o máximo de realismo possível, o problema da definição de um critério que permitisse uma correção homogênea de milhares de redações. As diretrizes que orientaram o estabelecimento do critério foram, posteriormente, divulgadas, merecendo destaque o seguinte aspecto:

> Buscou-se... operacionalizar um critério que atenda, a um tempo, a dois pressupostos básicos: uniformidade do comportamento analítico e interpretativo' dos diferentes avaliadores, por meio de um padrão objetivo de avaliação; exequibilidade, pela discriminação de um mínimo de subitens que permita considerar um máximo de aspectos determinantes da nota global. (1977a)

O interesse maior dos responsáveis pela elaboração do critério centralizou-se na homogeneidade dos julgamentos. O problema é de importância relativa no caso das redações realizadas em sala de aula, no decorrer de um curso, quando as flutuações de julgamento de um professor se compensam e não prejudicam, ao término da sequência instrucional, a avaliação global do estudante. A situação, no contexto do vestibular, é diferente: a população de candidatos é desconhecida, com escolaridade bem diversificada, e o número de redações a corrigir, em curto espaço de tempo, é elevado. Legítima é, portanto, a preocupação com o estabelecimento de parâmetros que possibilitem avaliações uniformes.

## CRITÉRIO DE CORREÇÃO – UM RESUMO

Elaborado por uma equipe de três professores universitários e um professor do 2º Ciclo, do Ensino Oficial do Estado de São Paulo, todos com experiência no ensino de nível médio, o critério foi aplicado, em janeiro de 1976, por quarenta professores, que ofereceram, ao término dos trabalhos, sugestões para modificá-lo em alguns de seus aspectos. Examinadas as críticas e propostas de modificação, a equipe elaborou nova versão do critério, divulgada em 1977, na forma seguinte:

**A correção da dissertação deverá considerar os aspectos abaixo discriminados:**

| | Valores máximos |
|---|---|
| 1 - **ESTRUTURA** - os julgadores verificarão se o trabalho apresentado pelo candidato é realmente uma dissertação e se essa dissertação constitui um conjunto articulado de partes em torno do tema proposto **(forma dissertativa, organicidade e unidade do texto)**; | 20 |
| 2 - **CONTEÚDO** - os julgadores verificarão se a dissertação apresenta ideias fundamentadas e coerentes, que demonstrem senso crítico e que possibilitem uma perfeita relação de entendimento entre o examinando e o avaliador **(elaboração crítica, coerência e clareza)**; | 30 |
| 3 - **EXPRESSÃO** - os julgadores verificarão se a dissertação apresenta: | |
| 3.1. adequação vocabular **(léxico)**; | 10 |
| 3.2. correção gramatical **(ortografia, morfologia, sintaxe, pontuação)**. | 40 |
| **Total** | **100** |

Ressalta do texto adiante transcrito, elaborado à guisa de apresentação, o cuidado dos responsáveis pela formulação do critério com a uniformidade e o rigor dos julgamentos.

> Embora não se devessem dissociar forma e conteúdo, ou seja, a articulação do pensamento e sua linguagem, é evidente que, na elaboração do critério, foi preciso fazer uma distinção convencional entre os elementos que entranhadamente constituem um texto. Assim é que, atribuindo valores parcelados a estrutura, conteúdo e expressão, o critério buscou fornecer aos avaliadores instrumentos mais

> precisos para o julgamento do texto como um todo. Os três itens são faces diversas do mesmo objeto, que é a prova total como se apresenta em sua redação final, resultante de um processo de elaboração em que o candidato deverá ter considerado necessariamente esses três aspectos. (1977a)

## APLICAÇÃO DO CRITÉRIO – PRIMEIRA EXPERIÊNCIA

Após a elaboração do critério, a equipe responsável aplicou-o na correção independente de quatro redações, antes de submetê-lo ao julgamento de outros avaliadores, ainda durante o treinamento. Os resultados dessa experiência inicial são apresentados na Tabela 1.

**TABELA 1 – Notas atribuídas a quatro redações por quatro examinadores, em correções independentes. Médias e desvios-padrão. Fundação Carlos Chagas - 1975**

| PROFESSORES | REDAÇÃO | | | |
|---|---|---|---|---|
| | A | B | C | D |
| 1 | 62,50 | 48,00 | 42,50 | 77,00 |
| 2 | 61,00 | 23,00 | 17,00 | 96,00 |
| 3 | 80,50 | 41,00 | 53,00 | 96,00 |
| 4 | 86,00 | 35,00 | 53,00 | 93,00 |
| $\ddot{X}$ | 72,50 | 36,75 | 41,37 | 90,50 |
| s | 10,67 | 9,17 | 15,14 | 7,88 |

Os elementos da Tabela 1 demonstram que houve divergências na aplicação do critério pela equipe responsável por sua elaboração, conforme indicam as dispersões (s) das várias notas. Analisando-se, entretanto, as posições atribuídas pelos quatro professores, observa-se que, na aplicação do critério, os julgadores foram unânimes em classificar as redações D e A como as melhores, nas posições 1 e 2, respectivamente. Verifica-se, também, quanto às provas B e C, que os julgadores concordaram que deveriam ocupar os postos 3 e 4, havendo dúvidas quanto à redação que se situaria na posição 3 ou 4, já que

dois deles acharam que a B deveria estar na posição 3, enquanto que os dois outros optaram pela redação C, para esse mesmo posto. Levando-se em consideração que a média dos julgadores é mais fidedigna do que um julgamento isolado (VIANNA, 1977), a aplicação do critério permitiu classificar as quatro provas, a partir da melhor, na seguinte ordenação: D, A, C, e B.

## APLICAÇÃO DO CRITÉRIO - NOVA EXPERIÊNCIA

A inclusão de uma prova de redação, no concurso vestibular de 1976, exigiu a seleção de quarenta (40) experientes professores para os trabalhos de avaliação. Os responsáveis pela coordenação das atividades de correção promoveram uma reunião de todos os envolvidos no processo, para apresentação e debates dos problemas. A preocupação da equipe coordenadora relacionava-se diretamente com a possível multiplicidade de julgamentos, pois a literatura técnica é rica em pesquisas que demonstram a diversidade dos critérios individuais de avaliação de provas dissertativas (VIANNA, 1976). A reunião objetivou, sobretudo, sensibilizar os avaliadores para os problemas da correção e vantagens da adoção de um único critério, a fim de atenuar as conhecidas diferenças individuais de julgamento.

Um documento explicativo, detalhando o critério (VIANNA, 1977 b) foi divulgado para encaminhar as discussões e detalhes. A parte introdutória deste documento mereceu especial atenção, sendo enfatizada, mais uma vez, a importância da homogeneidade dos julgamentos.

> A correção de redações é tarefa complexa, que envolve grande quantidade de variáveis. O critério geral... procurou abranger os aspectos que devem ser necessariamente observados, a fim de que se obtenham os dados fundamentais para uma avaliação um tanto quanto possível objetiva."
>
> "Sempre que lemos um texto, é inevitável sermos envolvidos por uma impressão geral, favorável ou desfavorável. Todo professor passa por essa experiência, mesmo que não o queira. No caso presente, o fato de os candidatos estarem empenhados numa situação competitiva obriga-nos a um maior controle da subjetividade. O que se pretende é o estabelecimento de critérios que possibilitem uniformidade

de julgamento por parte de um grupo de avaliação marcado pela diversidade de seus elementos.

Após os debates e a aceitação do critério pelo grupo de avaliação, distribuíram-se as redações anteriormente avaliadas pela equipe de coordenação, solicitando-se aos professores que a corrigissem segundo o critério. A redação D, por solicitação dos examinadores, foi avaliada com base na impressão geral. A Tabela 2 apresenta os resultados dessas correções.

**TABELA 2 – Notas atribuídas por quarenta professores a quatro redações, em correções independentes. Médias e desvio-padrão. Fundação Carlos Chagas - 1976**

| NOTAS | REDAÇÕES | | | |
|---|---|---|---|---|
| | FREQUÊNCIAS | | | |
| | A | B | C | D |
| 90-96 | -- | -- | -- | 8 |
| 83-89 | 1 | -- | -- | 10 |
| 76-82 | 3 | -- | -- | 6(*) |
| 69-75 | 1 | -- | -- | 13 |
| 62-68 | 9 | -- | 2 | 1 |
| 55-61 | 9(*) | -- | 3 | 2 |
| 48-54 | 5 | 1 | 5 | -- |
| 41-47 | 4 | 3 | 2 | -- |
| 34-40 | 6 | 8 | 16(*) | -- |
| 27-33 | 1 | 11(*) | 7 | -- |
| 20-26 | 1 | 12 | 2 | -- |
| 13-19 | -- | 5 | 2 | -- |
| 6-12 | -- | -- | 1 | -- |
| Notas | 40 | 40 | 40 | 40 |
| Ẍ | 55,52 | 29 | 38,57 | 79,37 |
| s | 14,18 | 9,27 | 12,45 | 8,97 |

(*) Assinala a classe em que se localiza a média

A análise da Tabela 2 possibilita observar, inicialmente, que as médias do grupo de quarenta professores ordenaram as quatro redações nas mesmas posições apresentadas pela equipe de coordenação. Ainda que as posições sejam idênticas, ressaltam, no caso, as diferenças entre os dois conjuntos de médias. A primeira correção, sob responsabilidade da equipe que elaborou os critérios, foi bem mais tolerante que a de quarenta professores, cujos resultados, entretanto, tendo em vista a teoria da fidedignidade, são mais precisos, por traduzirem um número bem maior de apreciações.

Essas diferenças entre as médias sofreram a influência da grande amplitude das notas dos quarenta professores, conforme a Tabela 3.

**TABELA 3 – Amplitude das notas atribuídas a quatro redações, em correções independentes, por quarenta professores. Fundação Carlos Chagas - 1976**

| NOTAS | REDAÇÕES | | | |
|---|---|---|---|---|
| | A | B | C | D |
| Mínima | 88 | 49 | 62 | 95 |
| Máxima | 25 | 14 | 6 | 60 |
| Amplitude | 64 | 36 | 57 | 36 |

A essa altura, configuram-se de imediatos dois problemas graves que, em princípio, podem ter implicações no contexto do vestibular:

**1º)** a grande amplitude das notas, ao que tudo indica, refletiria o fato de que os professores não estariam aplicando os mesmos padrões de correção estabelecidos pelo critério. A atribuição das notas, em muitos casos, foi aparentemente pouco cuidadosa, exemplificados pela redação C, cujas notas variaram de 62 a 6, com uma amplitude, portanto, de 57 pontos; e pela redação A, cujas notas tiveram uma amplitude de 64 pontos;

**2º)** outra constatação, com base na Tabela 2, refere-se à sujeição do estudante à equação pessoal do examinador: seu êxito ou fracasso vai depender da sorte de ter a sua prova corrigida por este ou por aquele professor.

A Tabela 4 reflete uma situação hipotética: considerou-se que somente uma nota superior a 40, nas distribuições apresentadas na tabela 2, traduziria aprovação. A seguir, calculou-se, percentualmente, quantos professores aprovariam ou reprovariam cada uma das quatro redações.

**TABELA 4 – Porcentagem de professores que aprovaram e dos que reprovaram, com base em notas atribuídas a quatro redações, em correções independentes, admitindo-se para aprovação um desempenho superior a 40. Fundação Carlos Chagas - 1978**

| REDAÇÃO | PORCENTAGEM DE | | Total |
|---|---|---|---|
| | Aprovação | Reprovação | |
| A | 80 | 20 | 100 |
| B | 10 | 90 | 100 |
| C | 30 | 70 | 100 |
| D | 100 | - | 100 |

A correção de uma redação não é assunto pacífico, conforme mostram as porcentagens da Tabela 4. Alguns consideram o trabalho digno de merecer aprovação; outros, ao contrário, o julgam destituído de valor. Os dados sintetizados nas Tabelas 2, 3 e 4 põem em destaque a possível influência do professor no êxito ou insucesso do aluno.

## APLICAÇÃO DO CRITÉRIO – VERIFICAÇÃO DE UMA HIPÓTESE

As atividades práticas de 1976 proporcionaram elementos para a reformulação do critério e o estabelecimento de novos procedimentos para sua aplicação. Aos avaliadores solicitaram-se relatórios individuais de sua experiência, especialmente das dificuldades encontradas na aplicação do critério, e apresentação de sugestões para a sua modificação, de modo a garantir o êxito de sua própria utilização. A partir do exame desses relatórios e de suas sugestões modificaram-se alguns elementos do critério, que passou a ter a versão anteriormente apresentada no presente estudo.

A dinâmica da reunião para debate do novo critério, por sugestão dos professores, sofreu, igualmente, modificações. Criaram-se grupos menores para permitir melhor interação dos avaliadores com a equipe de coordenação; desse modo, enquanto na experiência anterior o treinamento ocorreu com a presença de todos os avaliadores (40), nesta organizaram-se grupos com doze (12) professores apenas. Prepararam-se, também, dois documentos (1977a, 1977b) para exame mais detalhado: o primeiro estabelece a metodologia que orientou a elaboração do critério; o segundo define seus diferentes componentes e discute aspectos para a sua aplicação prática e quantificação dos trabalhos. A discussão desses documentos e a aplicação do critério a um novo conjunto de redações ocorreu em duas etapas, totalizando seis horas de atividades, nos meses de novembro e dezembro de 1977. Ao término dessas reuniões, havia, aparentemente, um consenso sobre a aplicabilidade do critério numa situação de vestibular.

Os exames de ingresso à universidade, em janeiro de 1978, permitiram testar a hipótese de que não haveria diferenças estatisticamente significantes entre as médias dos subconjuntos de redações corrigidos por professores que passaram pelas sessões de treinamento. Assim, 2.738 redações foram submetidas ao julgamento de 48 professores treinados, que, em média, corrigiram 57 redações num único dia, em dois períodos, num total de oito horas.

Após a correção das 2.738 provas, calculou-se, inicialmente, a média geral ($\bar{X}$ = 50,65); a seguir, a fim de atingir o objetivo maior do presente estudo, estabeleceram-se as médias de cada professor para cada um dos 48 subconjuntos. A Tabela 5 apresenta uma distribuição de frequência dessas médias, nos quarenta e oito subconjuntos. A amplitude das médias foi de trinta (30) pontos.

**TABELA 5 – Médias das notas atribuídas a 48 subconjuntos de redações corrigidos por diferentes professores submetidos ao mesmo treinamento. Fundação Carlos Chagas - 1978**

| X | F |
|---|---|
| 60-62 | 3 |
| 57-59 | 6 |
| 54-56 | 6 |
| 51-53 | 14 |
| 48-50 | 9 |
| 45-47 | 3 |
| 42-44 | 3 |
| 39-41 | 1 |
| 36-38 | 1 |
| 33-35 | 1 |
| 30-32 | 1 |
| N | 48 |

Observa-se, primeiramente, que 14 médias (29%) se acham na classe modal (51- 53), que também é a classe da média geral. A distribuição demonstra ter havido uma predileção pelos valores centrais, em torno da média. Assim, verificando-se as notas de cada prova, constatou-se que 2133 (77,90%) obtiveram notas na faixa de 30 a 70, numa escala de 0 a 100; que 331 provas (12,10%) receberam notas entre 75 e 100; e que, finalmente, 274 redações (10,00%) tiveram notas de 0 a 25. No conjunto de 2738 provas, atribuiu-se zero a 70, ou seja, 2,56% dos trabalhos receberam essa nota mínima. Apenas 3 provas (0,11%) mereceram nota plena (100). A Tabela 6 apresenta a distribuição de frequência das notas das 2738 redações.

TABELA 6 – Notas atribuídas a 2.738 provas de redação por 48 professores treinados. Fundação Carlos Chagas - 1978

| X | F |
|---|---|
| 91-100 | 19 |
| 81-90 | 111 |
| 71-80 | 201 |
| 61-70 | 396 |
| 51-60 | 512 |
| 41-50 | 631 |
| 31-40 | 460 |
| 21-30 | 232 |
| 11-20 | 93 |
| 1-10 | 13 |
| 0 | 70 |
| N | 2.738 |

A questão fundamental, anteriormente formulada na hipótese, consiste, portanto, em verificar se realmente existe diferença significante entre as médias dos quarenta e oito subconjuntos de provas. A expectativa, em que pesem elementos anteriormente apresentados, é de que não existam diferenças significantes, tendo em vista o cuidado do critério em eliminar todos os elementos capazes de gerar controvérsias. Além disso, a preocupação da equipe de coordenação em debater e analisar os vários problemas que poderiam ocorrer na fase de aplicação; a constituição de grupos homogêneos de avaliadores, no que tange à formação profissional e experiência docente; e, finalmente, a procura de consenso, supostamente obtido nas sessões destinadas à discussão e aplicação preliminar do critério, permitem esperar que não haja uma variação significante entre os examinadores.

A fim de testar a hipótese formulada, foi feita a análise da variância (ANOVA *one-way*) das médias das notas atribuídas pelos examinadores aos seus respectivos subconjuntos. A Tabela 7 apresenta os dados da ANOVA.

**TABELA 7 – Análise da variância das médias das notas atribuídas a 48 subconjuntos de redações corrigidos por diferentes professores submetidos ao mesmo treinamento. Fundação Carlos Chagas - 1978**

| FONTE DE VARIAÇÃO | GRAUS DE LIBERDADE | SOMA DOS QUADRADOS | QUADRADOS MÉDIOS | F |
|---|---|---|---|---|
| PROFESSOR | 47 | 108605,69 | 2310,76 | 7, 17*** |
| RESÍDUO | 2.690 | 866455,61 | 322,10 | |
| TOTAL | 2.737 | 975061,30 | | |

*** p ≤ 0,001 - Altamente significante

A análise da variância das médias dos 48 subconjuntos mostrou que existem diferenças significativas entre essas medidas de tendência central; consequentemente, não se pode admitir que essas discrepâncias sejam puramente casuais, devendo-se atribuí-las à falta de homogeneidade na utilização do critério pelos examinadores. Assim, a hipótese de que não haveria diferenças estatisticamente significantes entre as médias dos subconjuntos de redações, corrigidos por professores que receberam treinamento específico para esse fim, deve ser rejeitada; admite-se, desse modo, que os avaliadores, no processo de correção, empregaram critérios próprios, diferentes do proposto.

## CONCLUSÕES

Os elementos coletados no presente estudo possibilitam estabelecer as seguintes conclusões:

**1-** a aplicação de um critério de correção de redação pela própria equipe que o definiu não representa garantia de que haverá uniformidade nos resultados quantitativos apresentados;

**2-** ao corrigirem uma única prova, diferentes examinadores tendem a variações consideráveis nas notas, ainda que concordem quanto à posição dessa prova em relação às demais;

**3-** a variabilidade dos professores em relação a uma única prova é tão grande, mesmo supostamente usando um único critério, que a aprovação ou reprovação do estudante fica sujeita aos azares da sorte;

4- a utilização de um número considerável de examinadores, para fins de correção de redações, no contexto do vestibular, exige treinamento específico do pessoal docente, o que, todavia, não garante correção isenta de idiossincrasias individuais;

5- apesar da elaboração cuidadosa de um critério, da seleção de avaliadores entre profissionais altamente capacitados e da preocupação em estabelecer, por meio de treinamentos, normas de procedimento uniformes, as diferenças entre as médias dos avaliadores são estatisticamente significantes, conforme os dados da ANOVA – Tabela 7, inferindo-se, portanto, que os 48 professores participantes dessa experiência de avaliação utilizaram critérios – possivelmente, 48 critérios – diferentes daquele que foi apresentado, debatido e supostamente aceito por eles, na fase preliminar de treinamento.

## REFERÊNCIAS BIBLIOGRÁFICAS


FUNDAÇÃO CARLOS CHAGAS. *Apresentação do critério de correção*. São Paulo: Fundação Carlos Chagas, 1977a. Mimeo.

_____. *Explanação do critério adotado para julgamento da dissertação*. São Paulo: Fundação Carlos Chagas, 1977b. Mimeo.

VIANNA, Heraldo M. Redação e medida da expressão escrita: algumas contribuições da pesquisa educacional. *Cadernos de Pesquisa*, São Paulo, n. 16, p. 41-47, 1976.

_____. Flutuações de julgamento em provas de redação. *Cadernos de Pesquisa*, São Paulo, n. 19, p. 5-9, 1977.

# REFLEXÕES SOBRE A PRÁTICA AVALIATIVA

# AVALIANDO A AVALIAÇÃO: DA PRÁTICA À PESQUISA[1]

A análise das atuais práticas de avaliação nos diferentes níveis das escolas brasileiras mostra a necessidade da adoção de novas políticas e novos procedimentos para que seja possível o aprimoramento do processo de avaliação dos estudantes. É imperativo que se estabeleçam meios para verificar se diferentes tipos de aprendizagem estão sendo promovidos, sobretudo daqueles que exigem do aluno capacidades mais complexas. O problema não é de fácil solução para a nossa escola, especialmente em relação à prática avaliativa, por não existir concordância sobre questões bem imediatas, como os tipos de instrumento a construir e como chegar aos objetivos a que se propõe o sistema educacional.

Quando se considera, em termos apenas teóricos, as diversas formas de utilização dos instrumentos usados na avaliação educacional, constata-se que poderiam servir a diferentes propósitos:

1. avaliar a eficiência de professores, currículos, sistemas e programas educacionais;
2. identificar diferentes tendências quanto ao desempenho educacional;
3. determinar o progresso educacional do ponto de vista regional, nacional e até mesmo entre nações;

**1** Artigo publicado em *Estudos em Avaliação Educacional*, n. 5, p. 55-61, jan./jun. 1992. Trabalho apresentado ao XXIV Encontro Nacional da Fundação AMAE para Educação e Cultura no Instituto Granbery da Igreja Metodista, em Juiz de Fora, Minas Gerais (14-17.07.92).

4. possibilitar a definição e o planejamento de currículos, assim como a definição de novas políticas educacionais (NICKERSON, 1989).

A avaliação da eficiência de professores esbarra em interesses corporativistas e não faz parte da nossa tradição, que se limita a avaliar o aluno, assim mesmo de forma bastante precária. A avaliação de currículos, sistemas e programas começa a ser realizada, mas deforma restrita e limitada a algumas poucas experiências, como são o caso, no momento presente, da Avaliação da Jornada Única e das Escolas-padrão, em São Paulo, e da Avaliação do Ciclo Básico de Alfabetização (CBA), em Minas Gerais. A experiência mais vigorosa e com maior amplitude de Minas Gerais, que até o final de 1994 pretende avaliar todo o sistema educacional da rede oficial, gerando, assim, competências sobre esse tipo de avaliação, que ainda não possuímos, e concorrendo para a formação de uma cultura da avaliação, que ainda não temos.

A nossa carência atual em termos de avaliação decorre em grande parte da falta de continuidade nos trabalhos realizados, por intermédio de um processo de disseminação, que poderia servir de estímulo a outros empreendimentos na área de avaliação. A FUNBEC – Fundação Brasileira para o Ensino de Ciências, nos anos 60 e 70, realizou avaliação de currículos para a introdução de novas metodologias de ensino em Matemática, Física, Química, Biologia e Geociências, mas essa importante prática não teve continuidade em outras instituições, perdendo-se assim, parte do *know-how* adquirido com a colaboração de avaliadores de prestígio, como Hulda Grobman.

A experiência da Fundação Getúlio Vargas, no princípio dos anos 60, merece ser relembrada, tendo sido a primeira vez que se procurou construir um teste padronizado no campo da educação para avaliação de desempenho terminal. O projeto dirigido por Ruth Schaeffer e Nícia Maria Bessa, no Rio de Janeiro, criou um instrumento que seguia as linhas gerais do Iowa Basic Skills, e se destinava à avaliação de capacitações ao término do atual 1º ciclo. O projeto contou com a colaboração de figuras expressivas como Anne Anastasi, Frederick Davis e Robert L. Ebel, que participaram de treinamento para a formação de *expertise* entre educadores brasileiros. Houve um esforço no sentido de criar condições para o desenvolvimento de qualificações, mas, por diferentes razões,

o projeto não teve continuidade, restando dessa experiência apenas um manual de interpretação, elaborado com extremo cuidado metodológico, e uma pesquisa socioeconômica, que são modelos de trabalho científico da melhor qualidade em avaliação educacional.

A década de 70 apresentou grande interesse por avaliação de currículo, assistindo-se, no plano teórico, à divulgação do modelo sugerido por Stufflebeam – contexto, *input*, processo e produto (CIPP) –; no entanto, como é comum em nosso meio educacional, foi um momento transitório. Alguns trabalhos importantes, como os de Maria Amélia Azevedo e Clarilza Prado de Sousa, entre outros, foram realizados, mas também não tiveram continuidade.

Levantamento recente, a partir da revista *Cadernos de Pesquisa* (VIANNA, 1992), editada pela Fundação Carlos Chagas, mostrou que um amplo espectro de assuntos na área da avaliação foi analisado, discutido e pesquisado, restando, entretanto, uma indagação: – esses estudos chegaram ao conhecimento do professor e influenciaram na sua prática docente? A avaliação, lamentavelmente, não faz parte da formação dos docentes, quando muito é um tópico isolado, uma aula ou talvez uma unidade, mas não uma área de concentração.

Tentativas de avaliação do sistema educacional foram promovidas pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP/MEC), no final da década de 80, em pesquisas realizadas em escolas da rede oficial em 69 cidades do País, com o objetivo de identificar pontos críticos na aprendizagem de crianças do 1º grau (VIANNA; GATTI, 1988; VIANNA, 1989; VIANNA, 1989b), e, posteriormente, idêntica pesquisa do rendimento de alunos da rede privada foi concretizada. Houve um levantamento de importantes dados sobre o desempenho em Português, Matemática e Ciências de alunos da escola de 1º grau, mas não se possui um sistema adequado de disseminação de informações, que custam a chegar ao professor, quando realmente chegam, e nem sempre têm ressonância na escola. Ainda no início dos anos 90, o MEC/INEP procurou implantar um Sistema Nacional de Avaliação do Ensino Público de 1º grau, envolvendo custo-aluno, rendimento e gestão escolar.

Os anos 70 assistiram ao uso indiscriminado dos testes objetivos, por influência do processo de seleção para o ensino superior.

Elaborados sob o signo da improvisação, sem conhecimento de sua complexa tecnologia e dos seus fundamentos estatísticos, acabaram no descrédito. Ao longo desses anos foram ignorados os avanços da psicometria, área da estatística aplicada aos instrumentos de medidas educacionais e psicológicas que apenas tangencia os currículos para a formação de docentes. Seria inteiramente procedente, nesse momento, que se indagasse: para que servem os testes e/ou provas?

Os testes/provas (NICKERSON, 1989), quando considerados em relação aos alunos, podem servir para diferentes fins:

1. possibilitar o acesso aos vários níveis de escolaridade em diferentes escolas ou programas de ensino;
2. permitir a promoção em uma sequência educacional, nas suas várias fases, até a conclusão;
3. positivar deficiências a superar através de uma ação terapêutica;
4. identificar a possibilidade de acelerar (ou retardar) determinado programa;
5. orientar o processo instrucional por intermédio de uma avaliação contínua do desempenho, com a promoção de ajustamentos necessários à concretização da aprendizagem;
6. qualificar professores para o exercício de suas atividades docentes.

A análise dos instrumentos usados em nosso contexto educacional, muitas vezes construídos por instituições com excelente reputação, e, por isso mesmo, supostamente adequados às necessidades da avaliação, mostra elementos críticos, sobretudo no destaque de aspectos ligados à memorização e pouca ênfase no indicativo da capacidade de compreender e saber usar conhecimentos importantes em situações nóveis, revelando o real domínio do conhecimento graças à capacidade de aplicação.

A ausência de instrumentos capazes de medir compreensão e aplicação, objetivos maiores do planejamento educacional, mostra que os projetos de avaliação nem sempre têm condição de positivar se a escola – o sistema educacional – está realizando com sucesso, por intermédio do processo instrucional, um dos aspectos relevantes da tarefa a que se propõe. Isso gera de certa forma uma

cadeia de responsabilidades – os instrumentos não são capazes de determinar com êxito capacidades significativas, que a própria escola não as desenvolve conforme seria desejável. Assim, por via de consequência, a avaliação acaba prejudicada.

Existe toda uma cadeia de elementos inter–relacionados - os instrumentos de medida refletem a filosofia instrucional seguida em classe pelos professores e esses mesmos instrumentos (deficientes) determinam a forma pela qual os estudantes adquirem supostas capacidades. É necessário ressalvar que as avaliações são prejudicadas na medida em que os professores ensinam para o teste/prova, considerando que os seus resultados multas vezes são usados para determinar a eficiência do professor.

Ainda que compreensível do ponto de vista do professor, o ensinar para a prova acaba por comprometer o processo de avaliação no que tange à sua validade preditiva. É preciso levar em conta que não faz sentido preparar para o exame, tendo em vista os objetivos do processo instrucional. A ideia de ensinar para o teste, apesar de partir do pressuposto de que as provas determinariam o que o os professores ensinam e os alunos estudam, não é defensável, se for considerado que os instrumentos de avaliação nem sempre avaliam o relevante e desejável; desse modo, estaria sendo dada ênfase a atributos menores, em detrimento de capacidades mais importantes que, porém, não foram desenvolvidas face à relação ensino-teste-avaliação.

A questão de avaliar aquilo que é importante não é de solução fácil e imediata, tendo em vista certas constatações: os instrumentos para a medida de processos complexos são difíceis de construir, exigem pesquisas e investimentos, o que até agora não tem sido feito. A medida desses processos exigiria a avaliação de conceitos, princípios e relações, aspectos da aprendizagem nem sempre desenvolvidos em nossos currículos, que se preocupam com o domínio dos conhecimentos mais simples, quase sempre factuais e imediatos.

É preciso pensar e repensar o problema da avaliação no contexto brasileiro: avaliar é difícil, avaliar no campo educacional é extremamente difícil, e muitos não se dão conta das suas implicações pedagógicas e das amplas ressonâncias, inclusive no social e no econômico. A repetência, por exemplo, afeta a autoestima, e nem sempre é uma garantia de futuro êxito; ao contrário, as possibilidades de sucesso diminuem considera-

velmente. Essa é uma preocupação do atual Governo de Minas Gerais (1992), que em documento oficial declara haver todo um esforço no sentido de assegurar que o número de concluintes do ensino fundamental, com no máximo duas repetências, se eleve dos 15-18% atuais para 40% em cinco anos e para 60% em dez anos, quando, então Minas Gerais se terá equiparado ao México de hoje. A um custo médio de 220 dólares por aluno-ano, a repetência em Minas Gerais custa 110 milhões de dólares.

A psicologia educacional, nos dias fluentes apresenta grandes e novas concepções sobre aptidão, aprendizagem, desenvolvimento e rendimento escolar (SNOW, 1989; SNOW; LOHMAN, 1989.) Tudo isso está tendo amplas repercussões na área da avaliação educacional, especificamente nas pesquisas sobre validade de construto. Essa nova situação leva a crer que diferentes níveis e diversos modelos de avaliação devem ser adotados, sem abandono dos atuais, mas com um significado inteiramente diferente. O próprio professor precisa ser igualmente pesquisado, porque nada de importante se conseguirá sem a devida compreensão do seu papel e sem a sua irrestrita adesão aos trabalhos de avaliação.

A avaliação cognitiva é complexa e a sua complexidade vem aumentado na medida em que a psicologia cognitiva apresenta novos desafios, desvendando a multiplicidade de estratégias que levam a criança a diferentes tipos de aprendizagem. As crianças, sabe-se hoje em dia (SIEGLER, 1989), usam diferentes estratégias relativamente à soma, à subtração, à multiplicação, à soletração, à identificação de palavras, entre outros aspectos. A consequência disso é a valorização de um ensino voltado para as diferenças individuais e para uma Avaliação individualizada, do tipo formativo e baseada em critério e não em normas.

A avaliação em nosso contexto educacional acha-se na fase da pré-história, preocupada com problemas menores e sem significado efetivo para o sucesso de uma aprendizagem realmente consequente para o ser humano diante dos desafios que a sociedade constantemente apresenta. Assim, como avaliar as diferentes estratégias usadas na área cognitiva? A indagação está apresentada, mas dificilmente ter-se-á uma resposta satisfatória e imediata. É um *challenge* que devemos aceitar, esperando que algum dia se possa chegar a uma solução definitiva para o problema, que possivelmente será obtida por meios não convencionais,

em relação às atuais práticas de avaliação.

A avaliação procura entre seus objetivos verificar a capacidade de pensar criticamente, sem se dar conta de que somente se mede e avalia o que se pode definir operacionalmente, o que toma difícil a mensuração desse construto, face ao caráter extremamente vago do conceito e à impossibilidade de fixar padrões de julgamento (NORRIS, 1989). O assunto, por sua importância, precisaria ser pesquisado a fim de aprofundar o problema e identificar formas de avaliar o pensamento crítico no contexto dos problemas da realidade, o que significa dizer que há necessidade da determinação de um número bastante considerável de variáveis intervenientes nesse complexo problema.

É praticamente impossível discutir avaliação sem tratar, necessariamente, do problema da validade (NORRIS, 1989; FREDERICKSEN; COLLINS, 1989). Até que ponto a preocupação com os múltiplos problemas da validade chegou até nós é um questionamento que deveria ser objeto da reflexão de todos os avaliadores, inclusive de professores. Os testes aplicados em sala de aula são realmente válidos? Os instrumentos usados em pesquisas educacionais apresentam comprovada validade? Quais as evidências empíricas de que os resultados obtidos são realmente válidos? O que validar: o instrumento, os resultados ou a ideia que constitui a infraestrutura de todo o processo de avaliação? Quase sempre avaliamos construtos: rendimento escolar, capacidade de expressão escrita etc., mas qual a prova que se possui da validade de construto? É preciso atentar para as consequências de uma avaliação realizada sem base em uma fundamentação teórica (*rationale*) e sem evidencias empíricas que sustentem possíveis inferências e decisões com seus múltiplos resultados práticos, às vezes irreversíveis.

Os diversos problemas da avaliação precisam ser pesquisados para que a avaliação realizada na escola venha a ter um papel importante no processo ensino-aprendizagem, no qual está integrada. Somente a pesquisa sobre avaliação e a prática constante da avaliação serão capazes de criar uma cultura da avaliação e dar credibilidade à avaliação no contexto das atividades educacionais, sujeitas a constantes desafios. A informática, por exemplo, começa a ser introduzida em nossas escolas de maneira ainda incipiente, mas o uso de computadores na área educacional significará um novo repto à avaliação, que se

verá diante da possibilidade de poder avaliar como as pessoas compreendem e pensam (FREDERICKSEN; COLLINS, 1989); contudo, ainda temos um longo caminho a percorrer, necessitamos formar novas competências, a fim de enfrentar os problemas em benefício de um ensino eficiente em nossas escolas, sobretudo as responsáveis pela educação fundamental.

## BIBLIOGRAFIA


FREDERICKSEN, J. R.; COLLINS, A. A Systems approach to educational testing. *Educational Researcher*, v. 18, n. 9, p. 27-32.1989.

NICKERSON, R. S. New directions in educational assessment. *Educational Researcher*, v. 18, n. 9, p. 3-7, 1989.

NORRIS, S.P. Can we rest validity for critical thinking? *Educational Researcher*, v. 18, n. 9, p. 21-26, 1989.

SIEGLER, R. S. Strategy diversity and cognitive assessment. *Educational Researcher*, v. 18, n. 9, p. 15-20, 1989.

SNOW, R. E. Toward Assessment of Cognitive and Conative Structures in learning. *Educational Researcher*, v. 18, n. 9, p. 8-14, 1989.

SNOW, R. E.; LOHMAN, D. F. Implications of cognitive psychology for educational measurement, In: LINN, R. L. (Ed.). *Educational measurement*. 3th ed. New York: Macmillan, 1989. p. 263-331.

VIANNA, Heraldo M. Avaliação do rendimento de alunos de escolas do 1º grau da rede pública: um estudo em 20 cidades. *Educação e Seleção*, São Paulo, n. 19, p. 33-98, jan./jun. 1989a.

_____ . Avaliação do rendimento de alunos de escolas de 1º grau da rede pública: um estudo em 39 cidades. *Educação e Seleção*, São Paulo, n. 20, p. 5-56, 1989b.

_____. Avaliação Educacional nos *Cadernos de Pesquisa*. Cadernos de Pesquisa, São Paulo, n. 80, p. 100-105,1992.

VIANNA, Heraldo M.; GATTI, Bernardete A. Avaliação do rendimento de alunos de escolas de 1º grau da rede pública: uma aplicação experimental em 10 cidades. *Educação e Seleção*, São Paulo, n. 17, p. 5-52, jan./jun. 1988.

# A PRÁTICA DA AVALIAÇÃO EDUCACIONAL: ALGUMAS COLOCAÇÕES METODOLÓGICAS[1]

A avaliação educacional nos dias de hoje atrai a atenção de pesquisadores face à necessidade de promover o levantamento de um amplo espectro de elementos que possibilitem a análise dos sistemas educacionais em suas várias dimensões. Esse processo, entretanto, atravessa momento crítico, porque, conforme Stufflebeam *et al.* (1971), não possui uma teoria perfeitamente estruturada que traduza um consenso entre educadores. Além do mais, inexiste uma tipologia de informações fundamentais para o processo decisório, há insuficiência de instrumentos e planejamentos adequados aos diversos fenômenos educacionais e, ainda, destaca-se a falta de um sistema que possibilite a organização, o processamento e o relatório de informações necessárias à avaliação educacional. O trabalho de avaliação exige, consequentemente, certa criatividade e independe da adoção de modelos desenvolvidos em outros contextos. Assim, a pesquisa em avaliação educacional, inclusive quando restrita ao rendimento escolar, exige definições e procedimentos *ad hoc*, para que possa ser realizada com pleno êxito. É evidente que, não havendo um consenso difícil, às vezes, entre educadores ocorram, especialmente nessa área, incompreensões quase

**1** Artigo publicado em *Cadernos de Pesquisa*, n. 69, p. 40-47, maio 1989.

sempre centradas em aspectos não necessariamente relevantes para os propósitos da avaliação.

É necessário ressaltar que, ao contrário da pesquisa, a avaliação educacional não visa à generalização dos resultados com vistas ao estabelecimento de princípios ou de leis. A avaliação tem por objetivo gerar conhecimentos que levem a decisões que tenham consequências imediatas na prática educacional. Desse modo, no caso particular do rendimento escolar, por exemplo, a avaliação procura obter dados que possibilitem a tomada de decisões relativas ao desempenho individual e/ou coletivo em face de um determinado programa curricular. A partir de elementos empiricamente coletados, a avaliação procura descrever, da melhor forma possível, o fenômeno considerado, para possibilitar a fundamentação do processo decisório com base em dados da realidade.

A avaliação educacional, como área de investigação científica é recente, e as posições teóricas sobre as prioridades a considerar contribuíram para a formulação de diferentes definições e a criação de grande variedade de modelos. O problema primeiro está em definir, com relativa precisão, duas palavras frequentemente usadas de forma intercambiável: medir e avaliar.

Medir é uma operação de quantificação, em que se atribuem valores numéricos, segundo critérios preestabelecidos, a características dos indivíduos, para estabelecer o quanto possuem das mesmas. O índice quantitativo, obtido por intermédio da medida, identifica o status do indivíduo face à característica. Com referência à avaliação, a medida é um passo inicial, bastante importante, às vezes, mas não é condição necessária, nem suficiente, para que a avaliação se efetue. Eventualmente, a medida pode levar à avaliação que, entretanto, só se realiza quando são expressos julgamentos de valor.

Avaliar é determinar o valor de alguma coisa para um determinado fim. A avaliação educacional visa, pois, à coleta de informações para julgar o valor de um programa, produto, procedimento ou objetivo (WORTHEN E SANDERS, 1973) ou, ainda, a apreciar a utilidade potencial de abordagens alternativas para atingir determinados propósitos. A avaliação refere-se, assim, a atividades sistemáticas ou formais para o estabelecimento do valor de fenômenos educacionais (POPHAM, 1975), quaisquer que sejam.

A avaliação, para alguns, é um processo assistemático baseado na opinião de um especialista, é um julgamento emitido por um profissional. São comuns, na área educacional, "avaliações" informais para a tomada de certas decisões. Um livro é adotado em vez de outro, uma metodologia de ensino é empregada em substituição à outra, apenas com base em avaliações assistemáticas e impressionistas. A chamada "avaliação", nesses casos, limita-se a uma escolha, com base em percepções, da que seria a melhor alternativa. É, pois, uma simples opção, sem fundamento científico. A avaliação, ao contrário, decorre de um esforço sistemático para definição de critérios, em função dos quais se coletam informações precisas para julgar o valor de cada alternativa apresentada. Avaliar é, assim, emitir um julgamento de valor sobre a característica focalizada, podendo esse valor basear-se parcial, mas não exclusivamente, em dados quantitativos.

A Tyler (1942) coube a difusão da definição de avaliação como um processo de comparação entre os dados do desempenho e os objetivos instrucionais preestabelecidos. Essa definição desfruta de grande receptividade nos meios educacionais e, com pequenas variações, foi incorporada a alguns modelos teóricos como, por exemplo, o de Hammond (s.d.) e o de Metfessel e Michael (1967).

Stufflebeam *et al.* (1971) desenvolveram um modelo centralizado na ideia de que a avaliação deve permitir aos administradores a tomada de decisões e, coerentemente, definiram avaliação como o processo de identificar e coletar informações que permitam decidir entre várias alternativas.

Outros teóricos, juntamente com as suas contribuições de estratégias para a investigação avaliativa, propuseram também definições que, em maior ou menor grau, são aceitas por muitos praticantes da avaliação educacional. Entre essas definições, destacam-se a de Provus (1971), que apresenta a avaliação como um processo de comparação entre desempenho e padrões, e a de Stake (1967), que a caracteriza como descrição e julgamento de programas educacionais.

A avaliação, como campo emergente na área educacional, tem recebido contribuições provenientes de várias fontes, entre as quais destacam-se as de Michael Scriven, que marcaram, profundamente, a teoria da avaliação educacional. Scriven (1967) concebe esse processo como um levantamento sistemático de informações e sua posterior análise para fins de determinar o

valor de um fenômeno educacional. Essa definição, centralizada no problema do valor, influenciou o pensamento de grande parte dos teóricos e, praticamente, dos estudiosos da avaliação moderna, inclusive de alguns que não se preocuparam em detalhar e explicitar a questão, como foi o caso de Stufflebeam. Analisada a conceituação estabelecida por esse último, verifica-se que ela também incluiu um julgamento de valor, ainda que não o tenha explicitado. Escolher essa ou aquela alternativa, isto é, decidir, conforme estabeleceu Stufflebeam, é julgar o valor de uma ou de outra alternativa, optando pela melhor. Assim, na definição de Stufflebeam *et al*. (1971), está implícito, também, um julgamento de valor (VIANNA, 1982a).

A avaliação no contexto educacional é uma necessidade imperativa e exige uma metodologia que possibilite a coleta de informações para decisões fundamentadas. Um esquema de planejamento frequentemente encontrado em projetos de avaliação é o baseado na análise das diferenças apresentadas antes e após o tratamento instrucional. Essa estratégia, ainda que útil em certas condições, nem sempre fornece informações detalhadas que permitam tomar decisões complexas. Outra opção estratégica, também amplamente utilizada em avaliação, é a do planejamento experimental, que caracteriza a pesquisa empírica, mas que nem sempre é suficientemente eficaz para a avaliação de alguns fenômenos educacionais, tendo em vista a circunstância de que a avaliação se processa num quadro natural, em que as situações nem sempre são bem estruturadas e, por isso, tornam-se difíceis as condições de controle exigidas pelo planejamento experimental. Assim, tendo em vista essa problemática, vários especialistas procuraram desenvolver novas estratégias para dar à avaliação um sentido mais eficaz.

As diferenças existentes entre os modelos decorrem do fato de estabelecerem prioridades diversas para os problemas de avaliação educacional. Assim, como exemplificação, sem aprofundar a análise de todos os modelos anteriormente mencionados, observa-se que Tyler (1942) se encontra na problemática da convergência entre desempenhos e objetivos instrucionais, Stake (1967) baseia-se na análise de variáveis antecedentes, intermediárias (*transactions*) e resultantes, Stufflebeam *et al.* (1971), através do exame do contexto, entrada (input), processo e produto, visa a obter informações que permitam a tomada de decisões pelos administradores.

O avaliador educacional, ao selecionar determinado modelo teórico, a fim de desenvolver um projeto, deverá levar em consideração a natureza do problema a investigar, os recursos disponíveis e sua própria situação pessoal. Os modelos não se propõem a resolver todos os problemas que se apresentem ao avaliador; objetivam, na verdade, permitir que o avaliador dimensione, adequadamente, os seus projetos, para evitar que deficiências de planejamento invalidem o processo e levem a falsas decisões (VIANNA, 1982a).

Cronbach, no artigo *Course improvement through evaluation* (1963), ainda que não apresente um modelo de avaliação, oferece um conjunto de ideias altamente provocadoras que tiveram grande impacto na década de 60, influenciando trabalhos como o de Scriven (1967) e o de Stake (1967) que, por sua vez, repercutiram profundamente na prática de avaliação educacional. O ensaio de Cronbach é de grande valor e discute, sobretudo, quatro pontos: a associação entre avaliação e processo de tomada de decisão, os diferentes papéis da avaliação educacional, o desempenho do estudante como critério de avaliação de cursos, algumas técnicas de medidas à disposição do avaliador educacional.

A avaliação, no seu sentido mais amplo, pode ser definida como um processo que visa à coleta e ao uso de informações que permitam tomar decisões sobre um programa educacional. A avaliação, portanto, segundo Cronbach (1963), deve ser entendida como uma entidade diversificada, que exige a tomada de diversos tipos de decisões e o uso de uma grande variedade de informações. A avaliação com vistas ao aprimoramento de currículos não deve ser confundida, como muitos o fazem, com a construção de instrumentos de medida e a obtenção de escores fidedignos, processos que, eventualmente, podem entrar no contexto da avaliação mas que não são indispensáveis para que ela possa atingir os seus objetivos.

Cronbach mostra que a avaliação é usada com o objetivo de tomar alguns tipos de decisões, entre os quais:

1. determinar se os métodos de ensino e o material instrucional utilizado no desenvolvimento de um programa são, realmente, eficientes;
2. identificar as necessidades dos alunos para possibilitar o planejamento da instrução, julgando o mérito dos es-

tudantes para fins de seleção e agrupamento e fazendo com que conheçam seu progresso e suas deficiências;

3. julgar a eficiência do sistema de ensino e dos professores;
4. Assim, de acordo com Cronbach, no primeiro caso, a avaliação permitiria decisões que levariam ao aperfeiçoamento do currículo; no segundo, referirse-ia aos alunos submetidos a determinado programa; finalmente, no terceiro e último caso, as decisões seriam de natureza administrativa.

Cronbach discute, particularmente, alguns pontos fundamentais que não podem ser ignorados pelo avaliador educacional. Inicialmente, enfatiza o seguinte: "quando a avaliação visa ao aprimoramento de cursos, seu principal objetivo é verificar quais os efeitos do curso, ou seja, quais as mudanças que produz no estudante". O problema, segundo a perspectiva desse estudioso, não está em determinar se um curso é eficiente ou ineficiente apenas. É preciso lembrar que os resultados da instrução são multidimensionais e, desse modo, a avaliação deve promover o mapeamento de todos os efeitos do curso em cada uma de suas dimensões. Um erro frequente está na concentração em um único escore de diversos desempenhos esperados após a realização de um curso. Isso pode ser enganador, pois um insucesso numa dimensão pode ser compensado pelo sucesso em outra. Os escores compostos englobam e, muitas vezes, ocultam julgamentos sobre a importância de vários resultados; desse modo, para fins de avaliação, é importante que os resultados sejam apresentados separadamente, a fim de que se tenha uma ideia real das mudanças que estariam ocorrendo no estudante como decorrência da influência exercida pelo currículo.

"A avaliação presta um grande serviço quando identifica os aspectos do curso que necessitam de revisão", no dizer de Cronbach (1963). É evidente que qualquer especialista em currículo gostaria de apresentar evidências sobre a eficiência de seu produto; entretanto, conforme observa Cronbach, esses especialistas costumam relutar na aceitação de uma avaliação externa. O procedimento habitual consiste em submeter o produto à avaliação somente depois de terminado, com vistas a uma confirmação do que foi previamente estabelecido. Esse comporta-

mento não proporcionará bons resultados, além de traduzir um menosprezo pelo importante papel que o avaliador pode, efetivamente, desempenhar. A fim de apresentar um papel influente, a avaliação deve acompanhar o desenvolvimento do currículo, quando o mesmo ainda se acha em estado fluido, para usar palavras de Cronbach. A avaliação, assim entendida, possibilita a criação de conhecimentos sobre a natureza das capacidades que constituem os objetivos educacionais do projeto.

"A comparação de cursos não deve ser o objetivo dominante da avaliação", na concepção de Cronbach. A comparação de resultados de avaliações de cursos deve ser cautelosa, para evitar que decisões errôneas sejam tomadas. As diferenças entre os escores médios de diferentes cursos, geralmente, costumam ser pequenas, em virtude da grande diferença que há entre e intra grupos submetidos ao mesmo curso. A impossibilidade de equiparar (*equate*) diferentes grupos prejudica a interpretação dos resultados e representa um problema difícil de superar nos estudos comparativos de cursos. Além do mais, ainda conforme Cronbach, em experimentos educacionais, é difícil ocultar dos estudantes o fato de que integram um grupo experimental, sendo igualmente complexo o controle do viés dos professores numa situação experimental. Ocorrem mudanças comportamentais e, assim, nem sempre se pode afirmar, com convicção, que um determinado resultado decorre, efetivamente, da própria inovação ou do fato de alunos e professores terem sido colocados diante de uma situação experimental.

A comparação entre grupos pode oferecer resultados equívocos; desse modo, Cronbach (1963) propõe que estudos formais sejam planejados, sobretudo, para determinar o desempenho, após o curso, de um grupo perfeitamente conhecido, a fim de verificar objetivos importantes e ocorrência de efeitos colaterais. Cronbach chama a atenção para o fato de que, em um experimento onde os tratamentos comparados diferem em inúmeros aspectos, a ocorrência de uma pequena diferença numérica em favor da situação nova não significa grande coisa e não contribui para o aumento de nossos conhecimentos. Estudos analíticos, realizados em escala menor mas em condições controladas, de alternativas de um mesmo curso, oferecem melhores resultados do que pesquisas de campo que aplicam tratamentos dissimilares a grupos diferentes.

Cronbach destaca o fato de considerar mais importante os dados relativos a um item do que os escores do teste. O escore global pode dar confiança ou não em relação a um determinado currículo, mas pouco informa como aprimorá-lo. Outro aspecto importante em avaliação de currículo, conforme mostra Cronbach, refere-se à ajustagem do instrumento de medida ao currículo, que não deve ser motivo de preocupação para o avaliador, ainda que possa parecer surpreendente, porque, em outras situações, isso não deve ocorrer. Numa avaliação ideal de currículo, todos os tipos desejáveis de proficiência devem ser medidos e não apenas aqueles referentes a alguns objetivos selecionados pelo construtor do currículo. Se há interesse em saber se o currículo alcança os seus objetivos, impõe-se o ajustamento do teste a ele, mas, se a intenção é a de verificar até que ponto o currículo atende a interesses mais amplos, o ajustamento não é necessário, tendo em vista o desejo de que sejam verificados todos os objetivos possíveis.

Outro ponto enfocado por Cronbach diz respeito à distinção entre testes factuais e testes para verificar processos mentais complexos, segundo a terminologia de Bloom (1956). A classificação dos itens de acordo com as categorias propostas por Bloom é difícil e, às vezes, impossível. A classificação de uma resposta no nível de conhecimento ou raciocínio depende de como o aluno foi ensinado e não, apenas, da questão apresentada. Aplicar um teste somente para verificar se o aluno "sabe" ou "não sabe" um certo assunto não é inteiramente relevante para fins de avaliação de um curso, conforme Cronbach, importando, isto sim, medir o conhecimento em termos de profundidade, relacionamento e capacidade de aplicá-lo a novas situações.

A questão da especificidade das questões é considerada por Cronbach, que discute detalhadamente o problema. A especificidade, na maior parte das vezes, concentra-se no uso de uma terminologia própria do curso, que somente é compreendida por aqueles que tiveram a oportunidade de assistir a ele. Ainda que o conhecimento dessa terminologia seja importante, é mais importante, para fins de avaliação, a medida de compreensão de relação e de outras variáveis do curso, da que, em princípio, poderia ser verificada, também, em quem não foi diretamente submetido ao curso em questão.

O trabalho de Cronbach permite, entre outras, as seguintes conclusões:

1. A avaliação educacional requer a descrição de resultados; dessa forma, certas preocupações das medidas educacionais para a produção de escores precisos, visando a comparar indivíduos ou escores médios de diferentes cursos, pouco contribui para a avaliação educacional, cuja descrição dos resultados deve ser a mais ampla possível, ainda que às custas do sacrifício de uma suposta justiça e precisão.
2. A análise do desempenho em itens isolados ou em certos tipos de problemas fornece mais informações do que a análise de escores compósitos.
3. O objetivo da avaliação educacional não consiste em, simplesmente, aquilatar o valor de cursos, rejeitando-os ou aceitando-os, mas, sim, em ser uma parte fundamental no processo de desenvolvimento de currículos, através da coleta e do uso de dados que possibilitem uma compreensão mais profunda do processo educacional (VIANNA, 1982b).

A avaliação necessita considerar aspectos do meio educacional em que se desenvolverá, evitando, assim, a imposição de modelos nem sempre ajustáveis aos vários aspectos do sistema, ainda que bem estruturados do ponto de vista teórico. A realidade brasileira possui aspectos de grande singularidade, sendo preferível a adoção de algumas ideias gerais para orientação e fundamentação do trabalho de avaliação, como as que foram desenvolvidas por Cronbach (1963). Essas ideias são perfeitamente válidas em função de nosso contexto, mas não constituem, exatamente, um modelo. A adoção de um modelo transplantado de outras culturas pode criar uma situação artificial, sem desdobramentos práticos.

É preciso, no processo de desenvolvimento de uma avaliação, compreender o sentido de certos conceitos psicométricos, como validade, fidedignidade e discriminação, importantes sem dúvida em medidas educacionais, os quais, entretanto, devem ser ponderados em função de uma avaliação que não vise à seleção e ao prognóstico de comportamentos.

A validade de conteúdo é geralmente associada a uma amostragem representativa de conhecimentos e comportamentos adquiridos durante o processo educacional. A conceituação não é

pacífica, havendo quem exclua os comportamentos, limitando o problema à representatividade dos conteúdos programáticos. A questão dos comportamentos (objetivos instrucionais/educacionais) é, por sua vez, extremamente controversa. Há, inclusive, quem reaja ao ensino e à avaliação por objetivos. Sem entrar no mérito da questão, mas dentro da linha de considerar, na validade curricular, a simultaneidade dos conteúdos e dos objetivos, é preciso atentar para algumas particularidades desses últimos e para suas implicações no processo de avaliação. O fato de diversas taxonomias assinalarem níveis vários de objetivos não significa que, necessariamente, todos esses objetivos devam ser avaliados. Apenas aqueles que foram desenvolvidos no decorrer do processo instrucional devem ser, efetivamente, considerados. O desejável, às vezes, não corresponde à realidade, e um dos objetivos da avaliação é, justamente, descrever essa realidade tal como se revela por intermédio dos dados coletados. Ainda, com referência a comportamentos (objetivos), impõe-se atentar para o fato de que os mesmos são hierarquizados em ordem crescente de complexidade e abstração, sendo impossível, com bastante frequência, obter a concordância de especialistas sobre comportamentos mais heterogêneos, como, por exemplo, o de "avaliar" ou de "analisar". Desse modo, é comum, em avaliação educacional, a concentração de todos os comportamentos gerais e abstratos em uma única categoria: aplicação.

O fato de medir com fidedignidade não significa que se esteja realizando uma boa avaliação. A precisão com que se mede é condição necessária, mas não suficiente para a excelência de uma mensuração. Entre os atributos dos instrumentos de medida, importa realmente a validade, estando aí implícita a fidedignidade, não sendo verdadeira, entretanto, a situação recíproca. A fidedignidade, qualquer que seja a técnica usada para estabelecê-la (Alpha de Cronbach ou seu caso particular, as fórmulas de KuderRichardson), depende, fundamentalmente, da variância dos resultados dos escores. Quanto maior a variância, em princípio, maior será a precisão das medidas. Se esse aspecto é capital em uma avaliação somativa ou em um processo de seleção que, no fundo, é também uma avaliação de produto final, o mesmo não ocorre na avaliação formativa e/ou em uma investigação voltada para a avaliação de programas, a partir do desempenho escolar. O importante é a homogeneidade do grupo avaliado, que

refletiria domínio do material instrucional, entre outros aspectos, e não a heterogeneidade resultante da maior variabilidade entre os vários níveis de desempenho.

O cálculo de coeficientes de discriminação, expressos por diferentes tipos de medidas de correlação (r-bisserial, phi, entre outros), é presença obrigatória em estudos psicométricos na área da docimologia educacional, quando, especialmente, há necessidade de contrastar desempenhos extremos, a fim de verificar o comportamento dos itens/questões do teste/prova na separação dos indivíduos com vários níveis de domínio instrucional. A teoria clássica das medidas põe em destaque a importância desses coeficientes, favorecendo os itens/questões que discriminaram positivamente, ou seja, aqueles itens/ questões que foram respondidos, corretamente, pela maioria dos que obtiveram escores elevados e por poucos ou nenhum dos elementos situados no extremo inferior da distribuição de escores, em oposição aos itens/questões de discriminação negativa, isto é, que foram respondidos, em maior proporção, pelo grupo de elementos situados no extremo inferior da distribuição de escores e por pequena proporção dos situados no extremo superior. Os primeiros são valorizados por terem cumprido o seu papel discriminar; os segundos, condenados, pois não concretizaram o esperado discriminar. A avaliação formativa, assim como a avaliação com vistas a um determinado programa ou currículo, não objetiva discriminar, não pretende estabelecer diferentes níveis de desempenho. Ao contrário, a não discriminação, a homogeneidade do grupo, o máximo de respostas corretas, respeitadas as diferenças individuais, é o objetivo primordial de um ensino eficiente e de um programa ou currículo estruturado adequadamente, porquanto demonstraria sensibilidade ao processo instrucional e à sequência curricular. Ainda que possam constar de estudos de avaliação educacional, nem sempre os coeficientes de discriminação são peça essencial, apesar de necessários para análise de questões em uma avaliação somativa.

Um problema, muitas vezes, apresenta-se em avaliação educacional: é necessária a construção de testes padronizados e o uso de diferentes tipos de normas em investigações sobre o rendimento escolar? As palavras “testes padronizados” e “normas”, quase sempre, são usadas com imprecisão, con-

fundindo aplicação dos instrumentos de medida em condições padronizadas e segundo determinadas normas (orientações) com um teste planejado para proporcionar uma amostra sistemática do desempenho individual, aplicado segundo determinadas instruções, corrigido em conformidade com certas regras e interpretado por meio de informações normativas. A norma, por sua vez, é um conjunto de valores típicos descritivos do desempenho, num determinado teste, de um grupo específico de indivíduos, supostamente representativos de uma certa população. O contexto brasileiro, felizmente, não conhece os testes padronizados na área da educação. As restritas experiências realizadas no passado não prosperaram. Voltando à pergunta inicial, a resposta é negativa. A avaliação educacional pode, perfeitamente, prescindir de testes padronizados, que, além de extremamente caros, exigindo investimento de somas vultosas na sua construção, são demasiadamente genéricos e não possibilitam verificar o desenvolvimento programático realizado em sala pelo professor, no seu dia-adia. Os melhores instrumentos são as provas *ad hoc*, elaboradas por professores, com base na sua experiência, e revistas por outros professores, a partir de suas vivências pessoais. A avaliação educacional também não visa a desenvolver testes que sirvam *urbe et orbl*. Ainda que, teoricamente, isso seja possível, não faria sentido tal esforço e investimento, quando outros objetivos são prioritários em educação e a avaliação educacional pode contribuir para a sua concretização.

A avaliação educacional, segundo a conceituação apresentada, que a situou como um conjunto de operações 'visando levantar Informações que possibilitem uma tomada de decisões, como usar determinado livro, modificar um currículo, identificar deficiências no processo de aprendizagem, positivar falta de sensibilidade à instrução, pode adotar como amostra diferentes tipos de conjuntos: um pequeno grupo de alunos, uma série completa, alunos de um ou vários tipos de colégios, estudantes de diferentes áreas geográficas etc. A amostragem na avaliação educacional, assim como em outros tipos de investigação na área da educação, sofre limitações decorrentes do tempo, de pessoal e, especialmente, de orçamento. Todos esses fatores devem ser considerados. A questão tempo é, muitas vezes, a mais crucial, pois a tomada de decisões exige uma

coleta rápida para ação imediata, a fim de cortar a propagação de maiores danos em decorrência da demora da intervenção no processo educativo. A problemática de pessoal precisa ser considerada quando a natureza do trabalho exige material humano especializado, essencialmente na fase de aplicação dos instrumentos. A avaliação não é uma área de concentração, e poucos são os que, efetivamente, a ela se dedicam. A situação atinge contornos dramáticos quando os trabalhos de avaliação envolvem grandes regiões geográficas, às vezes de difícil acesso. A tudo isso acrescentam-se os custos da avaliação, que a tornam inviável em termos orçamentários, apesar de perfeitamente desenhada segundo as técnicas de amostragem. Assim, muitas vezes, seria desejável que todas as áreas geográficas estivessem adequadamente representadas com seus diferentes tipos de escolas, formando assim estratos nos quais, posteriormente, de forma aleatória, seriam selecionadas as instituições e indivíduos que, finalmente, integrariam a amostra. Nem sempre esse procedimento – que, supostamente, garantiria a representatividade da amostra – possibilita a factibilidade da avaliação em termos operacionais. Ao avaliador cabe decidir entre o ideal, fundamentado na matemática da amostragem mas irrealizável em termos concretos, e o possível em termos de realidade, que pode levar a inferências, a ilações, a implicações e, a partir dos dados levantados, chegar a conclusões que oportunizem decisões e modificações da realidade, objetivo final do processo de avaliação.

A adoção de um modelo matemático de amostragem seria necessária se a avaliação educacional, com o objetivo de identificar deficiências no rendimento escolar, por exemplo, visasse à generalização dos resultados a todo o sistema e não restringisse suas conclusões ao grupo de escolas e alunos participantes da amostra. Importante, no caso específico de uma avaliação educacional com vistas ao levantamento de "competências cognitivas" dos estudantes, é considerar a situação crítica das estatísticas sobre educação no Brasil. A estruturação de uma amostra para fins de avaliação educacional deveria "ter um conhecimento bastante seguro dos índices de repetência, promoção e evasão existentes no país", conforme Fletcher e Ribeiro (1988); contudo, é preciso convir que as discrepâncias entre as estatísticas educacionais coletadas por diferentes órgãos governamentais, em decorrência da diversidade das metodologias empregadas, são consideráveis,

chegando muitas vezes a 40%, como é o caso das estatísticas sobre população em idade escolar fora do sistema educacional. A questão da amostragem em avaliação educacional é complexa e, às vezes, difícil de solucionar, porque, malgrado o rigor dos procedimentos lógico-estatísticos, sempre subsiste a pergunta: a amostra selecionada é representativa do universo investigado?

O importante, em relação à atual estrutura de ensino em seus vários graus, é criar um sistema de avaliação externa que, de modo sistemático, informe aos responsáveis pelo ensino/educação os problemas da realidade pedagógica que ocorrem na escola e que se refletem nos diferentes níveis de capacitação cognitiva. A identificação de pontos críticos no desempenho escolar dos estudantes deve ser um dos objetivos do diagnóstico de deficiências, para que possam ser corrigidos os desvirtuamentos do processo ensino/aprendizagem.

O praticante da avaliação educacional, frequentemente, necessita posicionar-se em relação a problemas aparentemente menores mas que perturbam o seu trabalho, se não forem equacionados. As decisões sobre esses problemas são subjetivas na maioria das vezes. A vivência do avaliador e seu bom senso encontrarão uma solução para as situações que, imprevisivelmente, possam surgir. Assim, o avaliador, na construção dos instrumentos de mensuração, vê-se às voltas com a definição de um conteúdo programático que corresponda à média do que ocorre nas escolas do sistema, no caso de uma avaliação do rendimento. As Secretarias de Estado da Educação poderiam colaborar na solução do problema, mas, quase sempre, acham-se impossibilitadas, por não possuírem cópias dos programas desenvolvidos nas escolas. A legislação permite que cada escola estabeleça seus próprios conteúdos, havendo, portanto, grande diversidade de sequências curriculares em escolas de uma mesma cidade situadas próximas uma das outras. Algumas Secretarias possuem guias curriculares que oferecem uma certa ordenação ao conteúdo em suas linhas gerais, mas essa situação não é a regra comum. O avaliador vê-se compelido a criar um programa mínimo, não muito coerente às vezes, a partir dos aspectos comuns às várias programações. Uma solução alternativa e parcial para o problema estaria na análise dos livros didáticos utilizados, que se constituem em fonte subsidiária de informações sobre o que é desenvolvido em sala de aula pelos professores. O procedimento não

foge à prática ortodoxa da avaliação, sendo usado, inclusive, em outros contextos educacionais, para a construção dos chamados testes padronizados. Ajusta-se perfeitamente à nossa realidade, pois é sabido que, em muitas escolas, o livro didático é o verdadeiro "programa" curricular, como já foi comprovado em pesquisas nacionais. O livro didático, ainda que elemento colateral de informação, é um instrumento importante para que o avaliador possa caracterizar o que, efetivamente, ocorre no mundo da escola e, especialmente, em sala de aula.

A participação de professores na avaliação educacional, por intermédio da elaboração de questões e de objetivos instrucionais para uso futuro em investigações, constitui motivo de preocupação para alguns especialistas. O problema precisa ser dimensionado levando em consideração o tipo de avaliação e seus objetivos. Se a avaliação é interna, com finalidade formativa ou somativa, ninguém melhor do que o próprio professor para avaliar seus alunos, tendo em vista que foi ele quem estabeleceu objetivos, usou estratégias para conseguir o máximo de aprendizagem e orientou o aluno ao longo do processo instrucional. Se a avaliação é externa, usar professores da própria escola envolvida na investigação seria criar um novo problema, que desvirtuaria os resultados, levantando elementos que comprometeriam a tomada de decisão. As medidas realizadas seriam, sem sombra de dúvida, fidedignas, em grau talvez bastante elevado, mas a validade estaria irremediavelmente prejudicada, a não ser que se pudesse organizar um *pool* de professores igualmente capacitados, que gerariam um *pool* de questões de boa qualidade. Posteriormente, esses itens seriam selecionados de forma a constituir uma amostra representativa embasada na relação professores/alunos/escolas/currículos/objetivos/metodologias que integram a avaliação externa. A realidade educacional enfrentada pelos avaliadores é bem diversa do ideal, exigindo, assim, a adoção de práticas acauteladoras para o bom êxito do trabalho. A utilização de professores integrados no sistema educacional, independentemente da região ou área geográfica em que exerçam suas atividades, não contribui para desfigurar o processo de avaliação, desde que sua experiência resulte da prática objetiva do ensino, na realidade do dia a dia.

O problema do acesso às provas que serão aplicadas nas avaliações externas é conexo ao da elaboração de questões/obje-

tivos pelos professores das instituições avaliadas. Algumas experiências em avaliação educacional perderam o seu significado, justamente em face desse acesso. Ao tomar conhecimento das áreas de conteúdo objeto da avaliação, o professor, inconscientemente, tende a acentuar esses aspectos, provocando, portanto, um direcionamento favorável a seus alunos, em oposição aos demais sujeitos submetidos à avaliação. A avaliação externa, se não for bem compreendida pelos professores, gera uma certa ansiedade entre eles, em decorrência, possivelmente, do atual caráter punitivo de que, muitas vezes, reveste-se a avaliação em nosso contexto. O professor, desse modo, tende a criar uma imagem favorável para seu desempenho através do bom desempenho dos seus alunos. A avaliação procura obter dados para orientar o professor em seu trabalho, indicando as distorções no processo de aprendizagem. Busca, também, outros elementos que facilitem o processo decisório, não vindo a criar uma situação de constrangimento para o professor, a escola e o próprio sistema educacional.

A divulgação das provas usadas em pesquisas ligadas diretamente à avaliação costuma ser advogada por alguns elementos da comunidade educacional, criando uma situação de constrangimento para o avaliador que, a duras penas, construiu seu instrumental. Há uma tradição em nosso contexto social favorável à divulgação de provas e testes, inclusive por intermédio da imprensa, o que precisaria ser rompido. Ao avaliador cabe decidir o momento apropriado para divulgação do instrumental, no todo ou em parte. Alguns relatórios reproduzem fragmentos com tipos específicos de questões, outros se limitam à descrição geral do instrumental e, ainda, um terceiro grupo o divulga na sua inteireza, impossibilitando, desse modo, a reaplicação dos mesmos instrumentos em sucessivas avaliações. Ao longo do tempo, com a ampla divulgação dessas provas e a grande rigidez curricular, torna-se difícil a construção de novos instrumentos de medida sem o comprometimento da sua validade. Por outro lado, a disseminação de provas favorece o treinamento para a avaliação, como já vem ocorrendo em outros contextos que, tradicionalmente, usam testes padronizados. Ao avaliador cabe julgar, portanto, após considerar diferentes variáveis, o acesso às provas e a divulgação das mesmas. A prudência recomenda, entretanto, que o sigilo seja mantido,

se o objetivo for criar uma tradição relativa à avaliação externa, tornando-a parte integrante da vida escolar.

O uso de provas objetivas ou discursivas costuma ser apresentado como um problema para alguns avaliadores. A questão, em si, é um falso dilema, considerando-se que existe extensa literatura com base empírica, demonstrando, de modo insofismável, a adequação de ambos os tipos de prova para medir praticamente os mesmos traços, sendo, pois, inteiramente ociosa qualquer discussão sobre o assunto. Ao avaliador o problema deve se apresentar de uma outra forma, considerando a natureza da investigação e a população avaliada. Dependendo do contexto, talvez sejam recomendados instrumentos semiobjetivos apenas ou, então, inteiramente objetivos, em função do tempo disponível, da problemática relativa à sua aplicação e aos custos operacionais. A discussão sobre itens discursivos/objetivos não é bizantina, como poderia parecer à primeira vista, merece reflexão, mas não deve ser colocada em termos extremados, pois depende de um concurso de variáveis que interagem, cabendo assim ao avaliador tomar a decisão que melhor se ajuste ao quadro investigativo.

A avaliação, segundo a perspectiva ora desenvolvida, é um processo sistemático de levantamento de dados relativos a um determinado fenômeno, a fim de possibilitar a tomada de decisões com base em julgamentos de valor. A avaliação educacional aplicada à situação específica do rendimento escolar procura, desse modo, identificar as necessidades instrucionais dos alunos, com base em pontos críticos do seu desempenho em provas de escolaridade. Busca coletar elementos para fundamentar a análise da eficiência do sistema de ensino. A avaliação educacional está, basicamente, comprometida com a melhoria da qualidade do ensino e com a compreensão da influência dos professores no desenvolvimento de programas educacionais.

## REFERÊNCIAS BIBLIOGRÁFICAS

BLOOM, B. S. et al. *Taxonomy of educational objectives*: handbook 1; Cognitive domain. New York: David McKay, 1956.

CRONBACH, L. J. Course improvement through evaluation. *Teachers College Record*, New York, v. 64, n. 8, p. 672-683, May 1963.

FLETCHER, P. R.; RIBEIRO, S. C. *A educação na estatística nacional*. In: SEMINÁRIO DE AVALIAÇÃO DAS PNADS, 1980. Nova Friburgo: ABEP, 1988. Mimeo.

HAMMOND, D. L. *Evaluation at the local level*. Tuckson, Arizona: EPIC Evaluation Center, 1967.

METFESSEL, N. S.; MICHAEL, W. B. A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. *Educational and Psychological Measurement*, v. 27, 1967.

POPHAM, W. J. *Educational evaluation*. Englewood Cliffs (New Jersey): Prentice Hall, 1975.

PROVUS, M. M. *Discrepancy evaluation*. Berkeley: McCutcham, 1971.

SCRIVEN, Michael. The methodology of evaluation. In: STAKE, R. E. (ed.) *Curriculum evaluation.*. Chicago: Rand McNally, 1967. (AERA. Monograph, n.1).

STAKE, Robert E. The countenance of educational evaluation. *Teachers College Record*, v. 68, n. 7, p. 523-540. 1967

STUFFLEBEAM, D. H. et al. *Educational evaluation and decision making*. Itasca, Illinois, Peacock, 1971.

TYLER, R. W. General statement on evaluation. *Journal of Educational Research*, v. 35, p. 492-501, 1942.

VIANNA, Heraldo M. Avaliação educacional: algumas ideias precursoras. *Educação e Seleção*, São Paulo, n. 6, p. 63-70, 1982b.

_____. Avaliação educacional: problemas gerais e formação do avaliador. *Educação e Seleção*, São Paulo, n. 5, p. 9-14, 1982a.

WORTHEN, B. R.; SANDERS, J. R *Educational evaluation*: theory and practice. Washigton, DC: Charles A. Jones, 1973. 372 p.

# AVALIAÇÕES NACIONAIS EM LARGA ESCALA: ANÁLISES E PROPOSTAS[1]

*Rara felicidade de uma época em que se pode pensar o que se quer e dizer o que se pensa.*
TÁCITO, HISTÓRIAS

## INTRODUÇÃO E APRESENTAÇÃO DE PROBLEMAS

A avaliação educacional, especialmente a partir dos anos 90, passou a ser usada, no contexto brasileiro, em diferentes níveis administrativos, como tentativa de encontrar um caminho para a solução de alguns problemas educacionais mais prementes, esperando, possivelmente, que os processos avaliativos determinariam, entre outros resultados, a elevação dos padrões de desempenho, caso fossem conduzidos com o uso de tecnologias testadas na sua eficiência em outras experiências semelhantes, realizadas em diversos países, ainda que com culturas diferentes. Essa expectativa não se restringe unicamente ao âmbito nacional, sendo ocorrência bastante generalizada em quase todo o mundo ocidental, que concentra suas melhores esperanças nos resultados dessas avaliações.

**1** Artigo publicado em *Textos FCC*, v. 23, 2003. 41 p.

As avaliações apontam problemas, mas não os solucionam; outros caminhos deverão ser perseguidos.

A grande preocupação de educadores e de pessoas ligadas a problemas educacionais está na qualidade da educação, como demonstra o documento final da Conferência Mundial sobre Educação para Todos, ocorrida em Jomtien, Tailândia, em maio de 1990. O objetivo maior, na perspectiva oferecida no decorrer desse encontro, centrou-se na aquisição de conhecimentos, no desenvolvimento de habilidades e destrezas, na formação de atitudes, no despertar de interesses e na interiorização de valores; entretanto, não se considerou em que medida esses resultados se integrariam no contexto de uma sociedade em constante transformação, sujeita à intervenção de múltiplas variáveis nem sempre previsíveis.

É necessária uma reflexão sobre as avaliações ora operacionalizadas nos vários níveis do nosso sistema educacional, especialmente avaliações em larga escala, abrangendo a diversidade da nossa geografia multicultural, avaliações estas de natureza amostral e supostamente consideradas representativas em termos estatísticos. Fala-se, e com bastante destaque, ainda que nem sempre de forma consistente, na avaliação de competências e habilidades, mas de modo discutível e muito pouco consensual. Gostaríamos de invocar, neste ponto, antes de darmos prosseguimento às nossas reflexões, a citação de Tácito, em epígrafe, que David Hume usou na abertura de um de seus livros, deixando evidente, dessa forma, que os nossos comentários não visam a despertar susceptibilidades, mas tão somente a contribuir com a nossa reflexão para a análise de uma temática extremamente relevante no momento atual.

As questões que se impõem imediatamente, com o objetivo de aprofundar nossas percepções, podem ser propostas da seguinte forma: – são desenvolvidas competências e habilidades em nosso sistema educacional de uma forma sistemática, ou, explicitando, é o nosso ensino orientado para o desenvolvimento de competências? se for, qual a natureza dessas competências e supostas habilidades? Outra pergunta, que também reflete a nossa perplexidade: – se competências e habilidades foram promovidas, houve, efetivamente, preparo adequado dos educadores em relação a esse complexo e controvertido assunto? E quanto a atitudes, interesses e valores? As indagações

partem do princípio de que somente se pode avaliar aquilo que efetivamente foi desenvolvido, além de considerar que não se avalia em abstrato, mas considerando a problemática em que se situam os avaliados.

Quando pensamos em qualquer dos níveis da avaliação, micro ou macro, faz-se necessário que consideremos a complexidade do seu processo, que, ao longo dos anos, foi perdendo muito do seu caráter relacional aluno/professor, com vistas à orientação da aprendizagem, passando a concentrar-se, sobretudo, conforme chama atenção Kellaghan (2001), no desempenho institucional e no dos sistemas, como sucede igualmente em outras avaliações com objetivos mais amplos, de que são exemplos, no nosso caso particular, as avaliações promovidas na década de 90 pelo Governo Federal – SAEB – Sistema de Avaliação do Ensino Básico, ENEM – Exame Nacional do Ensino Médio, e ENC – Exame Nacional de Cursos.

Se forem considerados alguns aspectos dessas avaliações, constataremos que usam provas escritas, com questões objetivas e questões abertas, geralmente de resposta curta, havendo situações, entretanto, em que a prova de redação é exigida. Observamos, assim, que não existem provas práticas, orais ou avaliações observacionais, como lembra Kellaghan (2001), que seriam desejáveis para uma avaliação abrangente e conclusiva, mas impossível de se concretizar, somos forçados a reconhecer, em contextos que envolvem grandes massas, como no caso do ENEM/2002, por exemplo, que abrangeu quase 1,5 milhão de estudantes. Isso significa que não temos realmente um quadro avaliativo completo, que seja descritivo das diferentes dimensões do alunado, como seria desejável, mas uma simples métrica do que se supõe medir. É possível concluir, desse modo, que muitas competências e habilidades importantes no mundo atual não são efetivamente avaliadas, ficando implicitamente comprometida a definição do quadro educacional a ser configurado.

As avaliações são realizadas para diferentes fins, ainda segundo o posicionamento de Kellaghan (2001), destacando-se, inicialmente, como uma de suas prioridades, a identificação de problemas de aprendizagem, com o fito evidente de imediata superação do quadro apresentado. (Evitamos a palavra recuperação, tendo em vista o seu atual descrédito no meio educacional.) A realidade, entretanto, é bem diversa do imaginado e pretendido.

O impacto dos resultados pode ser considerado mínimo, por razões várias: – os relatórios, elaborados para administradores, técnicos e, em geral, para os responsáveis pela definição e implementação de políticas educacionais, não costumam chegar às mãos dos professores para fins de análise, discussão e estabelecimento de linhas de ação. São demasiadamente técnicos, empregando um linguajar pleno de tecnalidades muitas vezes desconhecidas dos docentes e que poderiam ser evitadas. Por outro lado, esses mesmos resultados são apresentados em termos globais, sem identificação, como seria desejável, das unidades escolares, referindo-se, quando muito, a unidades macro, os estados, e, nestes, eventualmente, às regiões geo-educacionais (superintendências ou delegacias de ensino). Ainda que os resultados dos desempenhos sejam apresentados em escalas elaboradas por intermédio de rigorosos procedimentos estatísticos, e com a especificação dos vários níveis correspondentes de competência, dificilmente os professores têm condições técnicas para interpretar dados que resultam da *expertise* técnica dos responsáveis pelos relatórios. Destaquemos, também, que há uma certa resistência, nem sempre explicitada, mas infundada, por parte de professores e alunos, aos resultados de avaliações amostrais, traduzindo, assim, certa dose de incredulidade em relação à generalização das conclusões. É comum ouvirmos:"a minha escola não fez parte da amostra" ou "os meus alunos não foram sorteados para a composição da amostra". Tudo isso faz com que importantes avaliações tenham o seu impacto, quando ocorre, bastante restrito, ou até mesmo seja inexistente, em relação ao sistema e a suas escolas.

Ao pensarmos nos problemas da avaliação, não nos podemos esquecer de que, assim como a motivação é fundamental para a aprendizagem, da mesma forma a motivação dos estudantes é importante para os trabalhos da avaliação. Entretanto, isso nem sempre ocorre e nem é objeto de consideração durante o seu processo. A avaliação é quase sempre impositiva, sem consulta a professores e muito menos a alunos. A avaliação, por sua vez, é igualmente repetitiva, no sentido de que, ao longo de vários semestres, os alunos fazem avaliações internas e externas, sendo que destas últimas não conhecem os resultados de seus desempenhos e das primeiras têm apenas um escore ou nota sem qualquer tipo de *feedback* que lhes possa servir de orienta-

ção. Esquecem-se as autoridades administrativas da educação e, às vezes, os próprios professores, que os alunos necessitam ser motivados para a avaliação, assim como, idealmente, são motivados para a aprendizagem, conforme destaque inicial. As avaliações, especialmente aquelas em larga escala, tornam-se monótonas, cansativas, geradoras de tensões e, muitas vezes, criadoras de conflitos, e como as avaliações não têm maiores consequências na vida dos avaliados, reagem os mesmos mecanicamente e respondem à *la diable* às várias questões apresentadas e, desse modo, as avaliações, reiteramos, perdem o seu significado, ainda que aos dados, resultantes de comportamentos inteiramente descompromissados, sejam aplicados procedimentos estatísticos complexos, que, por sua vez, geram todo um filosofar supostamente baseado em elementos considerados científicos e levam a decisões de repercussão, criando-se, assim, ideias falaciosas em grande parte da sociedade, que, apesar de tudo, passa a acreditar nas conclusões estabelecidas como se verdades absolutas fossem.

A avaliação – sempre considerando o caso brasileiro – procura, igualmente, estabelecer a eficiência dos sistemas, avaliando, indiretamente, o êxito da ação docente dos professores. Avaliar professores, direta ou indiretamente, é sempre um processo que demanda grande sensibilidade, pois gera múltiplas reações com ressonâncias negativas, qualquer que seja o contexto. A avaliação do professor, por sua vez, é vista com certa suspeita, pois, na concepção dos avaliados, e às vezes com justa razão, pode significar, em muitos casos, transferência de escola ou de cidade, redução salarial, diminuição do número de aulas, concessão de bônus para os supostamente melhores e, ainda, numa situação extrema, demissão. Tudo isso integra a mitologia educacional, bastante fértil em imaginar situações as mais diversas.

Avaliar o professor é sempre tarefa difícil e ingrata, mas deve ser feita, desde que com competência e, sobretudo, bom senso. A avaliação indireta, por meio do desempenho dos alunos, por sua vez, representa grande risco, com amplas consequências. É evidente que o processo ensino/aprendizagem se realiza por intermédio da interação professor/aluno, mas, por si, essa interação não resolve inteiramente a questão. Fatores externos à escola, inteiramente conhecidos pelos que transitam no mundo da pesquisa educacional, também têm importante papel no sucesso escolar, sendo suficiente citar alguns poucos como,

entre outros, a equivalência idade/série; horas de estudo no lar e a participação efetiva da família no acompanhamento das atividades escolares. O fracasso ou o baixo desempenho numa avaliação, portanto, nem sempre está relacionado ao professor, que, muitas vezes, por si, não tem condições de atuar visando à eliminação desses fatores. O ato de avaliar implica, necessariamente, considerar múltiplas variáveis, inclusive sociais, econômicas e culturais, que podem invalidar as ações subsequentes ao trabalho de avaliação.

Até que ponto as avaliações devem ser exclusivamente internas, eliminando-se a ocorrência de avaliações externas? Quando nos referimos a avaliações internas temos em mente as que são realizadas pelas escolas. É evidente que a avaliação na escola é parte do processo formativo, constituindo o trinômio ensino-aprendizagem-avaliação, sob orientação do professor. A avaliação interna pelos órgãos centrais do sistema é imprescindível, para fins de acompanhamento e reorientação dos procedimentos, se for o caso, além de constituir-se em fonte de desenvolvimento de competências e de apropriação de novas tecnologias por parte do pessoal do próprio sistema. As avaliações externas, realizadas quase sempre por proposta dos órgãos diretivos do sistema (Ministério da Educação; Secretarias de Estado da Educação), são recomendáveis, na medida em que representam um trabalho não comprometido com a administração educacional e as políticas que a orientam; são avaliações que traduzem uma visão de fora e supostamente isenta em relação a possíveis idiossincrasias próprias dos sistemas educacionais. Estas avaliações, entretanto, como será analisado mais adiante, representam um problema, quando abrangem regiões com grande amplitude de variação nas suas condições sociais, econômicas e culturais, face à ocorrência de possíveis comparações destituídas de sentido e a generalizações comprometidas, tendo em vista as diversidades apontadas que deveriam ser levadas em consideração na constituição de escores compósitos com valores agregados que traduziriam a maior ou menor influência da escola no desempenho educacional dos estudantes avaliados.

## ACESSO AO ENSINO SUPERIOR – UM QUADRO DISCUTÍVEL

Um aspecto a considerar, especialmente em relação às avaliações em larga escala, para fins de selecionar os melhores e mais capazes para o ensino superior, refere-se ao período de tempo em que são realizadas, sendo admissíveis duas situações: a avaliação ocorre de forma global, abrangendo alguns poucos dias seguidos; ou, então, em diferentes períodos, ao longo de vários semestres, no decorrer de três anos, em correspondência ao final de cada série do Ensino Médio, sendo esta modalidade bastante discutível. O primeiro modelo é seguido pela maior parte das instituições brasileiras de ensino superior, inclusive universidades e centro universitários. O período de tempo das avaliações quase nunca ultrapassa a quatro dias, mas num passado recente houve avaliações que duravam quase toda uma semana. Uma alternativa a esse tipo de avaliação, ora sendo executado por muitas instituições, consiste na avaliação em duas fases, sendo a primeira seletiva, com o objetivo de eliminar parte do grande número de candidatos ao ensino superior, e a segunda, classificatória, para atendimento do *numerus clausus* que regula o acesso por curso.

As avaliações anteriormente apresentadas, instituídas há mais de 90 anos, são altamente controversas, na formulação dos seus propósitos e no instrumental empregado. É um tipo de avaliação associada à problemática do alto número de sujeitos que terminam o ensino médio sem possibilidades do exercício de qualquer atividade profissional, restando-lhes a tentativa do acesso ao ensino de terceiro grau, que também tem graves problemas, mas com características específicas. É uma avaliação estressante e a qualidade dos instrumentos bastante comprometida, salvo em algumas universidades e fundações dedicadas especificamente à pesquisa e à avaliação, que desenvolveram e aprimoraram o seu *know-how* docimológico, inclusive usando complexas metodologias estatísticas para fins de análise de questões e da identificação de atributos psicométricos desejáveis; contudo, grosso modo, pode-se dizer que são avaliações *ad hoc*, com a construção reiterada, ano após ano – é um trabalho de Sísifo –, de novos instrumentos que nem sempre se revestem das características desejáveis, especialmente em relação à validade de conteúdo e à de predição, não havendo, também, preocupação maior com a fidedignidade (precisão) dos resultados,

que quase nunca é estimada, mas que, por intermédio de uma análise qualitativa crítica, pode ser inferida, considerando a não representatividade amostral dos conteúdos e das capacidades, e as deficiências técnicas na construção dos itens ou questões.

As avaliações em duas fases, uma seletiva e outra classificatória, no acesso ao ensino superior, inicialmente restrita a poucas instituições, hoje, entretanto, conta com maior número de adesões. A adoção desse modelo não resultou, salvo melhor juízo, de análises e considerações sobre a melhoria do processo; na verdade, procurou solucionar problema operacional, tendo em vista que, em muitos casos, há o envolvimento de centenas de milhares de estudantes. A segunda fase estabelece *a priori* como ponto de corte um valor igual, aproximadamente, a três vezes, em média, o número de vagas por curso, e com uma única avaliação, realizada por meio de um único instrumento voltado apenas para conhecimentos e algumas poucas capacidades, consegue reduzir a grande massa de sujeitos a um nível razoável, em termos econômicos, tendo em vista os custos operacionais das avaliações em larga escala. Estes selecionados passam, então, para a segunda fase classificatória. Uma situação extremamente bizarra se configura no caso, quando se relacionam os resultados das duas fases e são obtidos coeficientes elevados e positivos. Isso significa, primeiramente, que os melhores da segunda fase foram os igualmente melhores, em princípio, na fase inicial (seletiva), sendo a segunda fase, consequentemente, redundante, além de evidenciar a natureza repetitiva desta última fase.

Ao longo do Ensino Médio, em alguns casos, temos avaliações parceladas, ao fim de cada série, que, depois de terem seus resultados consolidados, geram um escore compósito que servirá para a fase classificatória do processo seletivo. Algumas poucas universidades, é bem verdade, seguem esse procedimento, reservando para os sujeitos submetidos a essa avaliação determinados percentuais de vagas. A "nova" sistemática, na visão de muitos, revestir-se-ia de maior racionalidade, evitando, inclusive, a chamada situação de stress de uma única avaliação; entretanto, é necessário atentar para o fato de que essa metodologia gera um desvirtuamento do Ensino Médio, que, supostamente, é dedicado à formação geral, mas, no caso presente, passa a ser inteiramente direcionado para o ensino superior, transformando-

-se em um curso meramente preparatório para o terceiro grau, e quanto ao stress, este acaba sendo triplicado ou, como colocou ilustre professor preocupado com problemas de ensino e repetência, o aluno ao invés de passar uma vez pela guilhotina, passa três vezes, sem maiores contemplações.

Ainda com relação à avaliação para acesso ao terceiro grau, e com apoio de órgãos do executivo e do legislativo estadual, começa a ser desenhado, sem maiores estudos e análises, e sem considerar suas numerosas implicações e sérios efeitos, um novo modelo de reserva de vagas – sistema de cotas – para estudantes oriundos do sistema público de ensino e estudantes negros, candidatos a instituições oficiais, na tentativa de superar um problema que na realidade se concentra na baixa qualidade do ensino fundamental e do ensino médio público, comprovada por pesquisas empíricas, inclusive muitas realizadas por órgãos oficiais. As primeiras novas experiências, nesse sentido, ocorreram no início de 2003, no Rio de Janeiro, rompendo, desse modo, o princípio da isonomia – igualdade de condições para todos – existente no sistema ora vigente de avaliação.

## AVALIAÇÕES SISTÊMICAS – ALGUMAS QUESTÕES CRUCIAIS

Ainda nos anos 90 houve grandes avaliações dos sistemas estaduais de ensino no Brasil, ligadas, na maioria das vezes, a projetos educacionais financiados pelo Banco Mundial. Essas avaliações apresentaram-se de diferentes formas: algumas, realizadas pelas próprias Secretarias de Educação; outras, por órgãos estaduais nem sempre diretamente ligados à área da educação; um terceiro grupo, com a colaboração de Fundações, instituições de direito privado especializadas na avaliação e seleção de recursos humanos; finalmente, um quarto grupo realizou suas avaliações sistêmicas estabelecendo consórcios com múltiplas instituições de ensino público e privado de terceiro grau, sob a coordenação de uma universidade de prestígio orientadora de todo o processo. Tudo isso gerou diferentes experiências, mas não contribuiu para a formação de um *know-how* coletivo, pois, na maioria dos casos, essas experiências não se transformaram em vivências que pudessem ser intercambiáveis e a própria

divulgação dos resultados foi precária, sem atender aos diversos segmentos educacionais potencialmente interessados nos resultados e nas conclusões das avaliações.

Algumas avaliações sistêmicas tiveram um caráter censitário, mas a maioria optou pela adoção de avaliações amostrais. As primeiras, ainda que apresentassem custos elevados, tendo em vista o número expressivo de alunos e a problemática de uma logística complexa, foi resultado de uma decisão política: – fazer com que todo o sistema participasse da problemática da avaliação e não se limitasse apenas a colaborar na aplicação dos instrumentos, mas fosse partícipe inclusive da construção dos instrumentos e dos trabalhos de uma correção preliminar nas respectivas escolas, discutindo, imediatamente, os primeiros problemas identificados e fossem antecipadas as primeiras providências para o seu saneamento, antes da divulgação dos resultados globais pelos órgãos centralizadores. Outros sistemas começaram com avaliações amostrais, que nem sempre tinham grande impacto, e evoluíram para avaliações censitárias, supostamente pelas razões anteriormente apontadas. A maioria, entretanto, optou por uma avaliação amostral, por representar economia de problemas operacionais e minimizar os custos, além de oferecer resultados igualmente confiáveis. As avaliações censitárias tinham a vantagem de apresentar os resultados por escola, município, Delegacia ou Superintendência de Ensino, e os dados globalizados por estado.

Observa-se nessas avaliações que o grau de sofisticação do tratamento estatístico dos dados variou grandemente. Inicialmente, houve uma tendência a apresentar os resultados de forma que fosse palatável para o sistema, que estivesse de acordo com a cultura educacional de todos os segmentos e seria ingenuidade imaginar que os professores do ensino fundamental ou do ensino médio tivessem suficiente conhecimento estatístico para entender práticas de análise supostamente novas, mas que já vigoravam em países mais avançados desde os anos sessenta, como é o caso da análise das questões por intermédio da metodologia da Teoria da Resposta ao Item (TRI). A impossibilidade de aplicação imediata dessas novas tecnologias decorreu, também, da inexistência de *hardware* nas Secretarias de Estado da Educação, que se utilizavam de outros órgãos, não necessariamente ligados à educação, para o processamento de dados, além, naturalmente, da falta de domínio na utilização dos pa-

cotes estatísticos com os novos procedimentos de análise.

A tendência atual que se observa, decorrido um decênio das primeiras avaliações sistêmicas, é a da opção por avaliações amostrais, seguindo as linhas gerais das grandes avaliações instituídas pelo Governo Federal, inclusive com o uso de questões integrantes do Banco de Dados do Instituto Nacional de Estudos e Pesquisas Educacionais – INEP – e já submetidas à pré-testagem. Naturalmente, a situação ao longo dos anos se alterou e nos dias fluentes as chamadas "novas" metodologias de análise são utilizadas com bastante frequência, ainda que o seu entendimento seja precário, tanto por parte do público mais diretamente interessado – a escola e os educadores –, como por muitos especialistas em avaliação que ainda não superaram os procedimentos canônicos em que foram formados, sobretudo os integrantes da geração que se formou nos anos sessenta, muitos dos quais optaram por abordagens qualitativas ou permaneceram identificados com a chamada Teoria Clássica das Medidas.

Outra questão observada nas primeiras avaliações relacionou-se ao tipo de instrumento a ser empregado, ocorrendo discussões se seriam instrumentos referenciados a critério ou referenciados a normas . O debate foi em termos da realidade nacional, que, inclusive, naquele momento, desconhecia os fundamentos desses dois tipos de instrumentos e, consequentemente, não tinha um domínio da sua tecnologia e da sua metodologia de análise. Ainda que ambos os tipos de instrumentos fossem viáveis para os fins desejados, prevaleceu o bom senso e a opção foi a de utilizar instrumentos referenciados a normas, mais adequado à tradição da nossa cultura pedagógica, que já o utilizava sem um conhecimento aprofundado dos seus fundamentos teóricos. Além do mais, nessas avaliações foi polêmica a consideração de que a mesma seria de natureza somativa, para usar a expressão de Michael Scriven, na sua obra clássica, *Methodology of Evaluation*. A discussão teve, entretanto, algum mérito. Foram realizadas palestras e cursos sobre avaliação por critério, mas esse novo tipo de instrumento passou a ser conhecido apenas por uma minoria de professores.

A avaliação por critério seria ideal para a avaliação de processo, para correção e superação de dificuldades de aprendizagem, mas esse tipo de avaliação ainda não foi incorporado à cultura nacional e deveria integrar o processo de educação con-

tinuada que se desenvolveu nos anos 90. Lamentavelmente, a chamada progressão continuada, impropriamente chamada de promoção automática, denominação que inclusive concorreu para o seu desvirtuamento, ainda não é bem aceita pela comunidade, apesar de esforços para esclarecimento da sua lógica e do seu significado, que pressupõem constante uso de diferentes tipos de trabalho avaliativo em todos os momentos do processo instrucional. Essa seria a ocasião apropriada para a introdução da avaliação referenciada a critério e aos trabalhos com grupos diversificados pelo mesmo professor, que muito teria a aprender com a prática das professoras nas escolas rurais, que trabalham simultaneamente com alunos que apresentam diferentes níveis de rendimento. Os professores deveriam ter treinamento específico, dispor de recursos e materiais didáticos para suprir possíveis deficiências dos grupos com características diferenciadas, mas nada disso ocorreu, criando-se, dessa forma, um certo confronto entre professores, alunos, comunidade e a progressão continuada, pela ausência de uma avaliação própria para atender a diversidade dos desempenhos.

A avaliação de sistemas durante os anos 90 e, sobretudo, no seu início apresentou um problema realmente crítico e somente parcialmente superado nos dias fluentes: – ausência de pessoal com formação específica em avaliação educacional, que, no contexto nacional, não é considerada área de concentração. Alguns problemas surgiram em decorrência dessa realidade, como as improvisações, em alguns casos, a subordinação aos chamados "especialistas", em outros, e a adoção de novas metodologias, sobretudo estatísticas, sem a posse do seu domínio, determinando, como decorrência, algumas situações verdadeiramente bizarras. Apesar de passado mais de um decênio do início das grandes avaliações, o problema ainda persiste e dificilmente será resolvido a curto prazo sem uma mudança de mentalidade e a criação de uma nova cultura educacional.

## SISTEMA DE AVALIAÇÃO DO ENSINO BÁSICO – SAEB

O Governo Federal, ao implantar um programa de avaliação abrangendo o ensino básico, o médio e o superior teve um gesto extremamente corajoso, considerando, entre outros as-

pectos, a amplitude da tarefa, a dificuldade na definição de padrões, os problemas técnicos nas decisões sobre os instrumentos e sua tecnologia, a possível subjetividade dos julgamentos de valor e a complexidade das operações logísticas. E chegamos, agora, a um ponto crítico em que se impõe a avaliação da própria avaliação (meta-avaliação) e, simultaneamente, a autoavaliação de seus procedimentos, para rever antigas ações e propor novas outras ações, à luz da experiência acumulada. A avaliação para aprimoramento do próprio projeto avaliativo é um imperativo a que não se pode escapar.

O Sistema de Avaliação do Ensino Básico – SAEB – é, sem sombra de dúvida, a nosso juízo, o melhor e o mais bem delineado dos projetos propostos pelo Ministério da Educação. Nele dever-se-ia concentrar todo o empenho governamental, por ser o ensino básico o fundamento para a construção do espírito de cidadania e o alicerce sobre o qual se apoiam os demais níveis educacionais; por isso, acreditamos que seus responsáveis se deveriam preocupar, particularmente, com duas das características dos instrumentos de medida voltados para o rendimento escolar, a validade de conteúdo e a validade consequencial.

A validade, segundo o consenso dos especialistas, não é uma característica geral, antes de tudo ela é específica. Um instrumento de medida não é válido em tese, pode ser válido para um curso, mas não para outro. Pode ser válido para um currículo, mas não para outro; para um professor, mas não para outro, inclusive, pode ser válido para uma escola, mas não o ser para outra instituição. A questão da validade é extremamente delicada em qualquer contexto educacional e, no nosso caso particular, precisamos considerar a formação da nossa nacionalidade, a grande diversidade social, econômica e cultural, demonstrada em todo o território brasileiro, que varia de regiões desenvolvidas, passando por zonas de transição e chega a imensas áreas com estruturas arcaicas. O problema da validade, reiteramos, precisa ser tratado com extrema cautela, a fim de evitar que a posterior análise dos dados possa levar a inferências destituídas de sentido. Tudo isso é um desafio, sendo forçoso atentar para a validade amostral ou de conteúdo dos instrumentos utilizados, para que sejam os dados representativos da diversidade da nossa geografia cultural. Os programas de pesquisa sobre o SAEB deveriam incluir, ne-

cessariamente, uma parte dedicada a estudos de validade, nas suas diferentes modalidades, evitando-se o tratamento tangencial da questão, como vem ocorrendo em alguns poucos trabalhos que discutem a problemática da avaliação.

Outro problema a considerar, no caso do SAEB, relaciona-se à validade consequencial, que se refere ao impacto da avaliação sobre o sistema, determinando mudanças de pensamento, gerando novos comportamentos, formando novas atitudes e promovendo novas ações. A validade consequencial reflete em que medida a avaliação faz realmente alguma diferença para a comunidade. Até agora a influência do SAEB, na nossa visão, tem sido bastante restrita na comunidade escolar, em que pese o sucesso jornalístico, com a publicação dos seus resultados nos vários órgãos da mídia.

O SAEB, ao divulgar o relatório de suas avaliações, apresenta a metodologia, os tratamentos a que foram submetidos os resultados e uma grande riqueza de dados e informações sobre os diferentes desempenhos; entretanto, esse documento, elaborado com extremo rigor técnico, acaba por se tornar inacessível à grande massa de interessados dentro e fora do campo da educação. A sociedade, por intermédio da publicação dos resultados em jornais, com inúmeros e bem construídos gráficos e tabelas, que procuram ser autoexplicativos, assiste a tudo sem entender bem o que se passa e, acreditamos, muitos pais se indagarão: – a escola do meu filho se saiu bem? o meu filho teve uma boa nota na avaliação? o meu filho foi melhor ou pior que os seus companheiros de classe? e os seus colegas de série se saíram melhor ou pior do que ele? São grandes incógnitas em uma situação pouco compreensível para a grande massa.

Queremos mais uma vez destacar a importância e o significado do SAEB, como avaliação de sistemas, mas é preciso que os responsáveis pela sua administração compreendam que diferentes setores da sociedade estão interessados em conhecer e discutir os dados do SAEB e a cada um desses segmentos deveria corresponder diferentes documentos, apresentados desde a sua forma mais completa, incluindo diferentes estatísticas, estudos de validade e análises dos vários desempenhos e suas capacidades, relatórios técnicos, enfim, até a sua versão mais simples, que poderia ser apenas um folder informativo, para divulgação entre os pais e demais integrantes da sociedade. Devemos con-

fessar, por ser de inteira justiça, que, em 2001, o INEP, compreendendo a relevância do problema ora exposto, promoveu em Curitiba, na Secretaria de Estado da Educação, uma reunião de elementos das outras Secretarias e pessoas ligadas à avaliação educacional para discutir a questão da disseminação do SAEB, ficando assentado que em 2002 apresentaria seus dados em relatórios com diferentes abordagens, para atender os vários segmentos da sociedade. Assim procedendo, e havendo a integração das escolas para discussões dos dados, acreditamos ser possível que, a médio prazo, talvez se possa começar a falar da validade consequencial do SAEB.

## EXAME NACIONAL DO ENSINO MÉDIO – ENEM – PROPOSTAS ALTERNATIVAS

A ideia de uma avaliação ao término do Ensino Médio provocou grandes expectativas em alguns ambientes educacionais, por corresponder a uma necessidade, considerando, entre outros aspectos, a expansão descontrolada da rede de ensino, especialmente no âmbito privado, que apresenta, como é do conhecimento geral, diferentes níveis, variando desde as escolas realmente excelentes, com elevado padrão de ensino, a escolas sem maiores compromissos. A criação de um Exame de Estado, ideia que surge recorrentemente, provoca grandes discussões, por ser uma medida bastante problemática, que acarretaria inúmeros e sérios problemas, sobretudo no atual quadro nacional. Felizmente, essa ideia não prosperou. Outros chegaram a falar na introdução de um exame semelhante ao *Baccalauréat* francês, o que poderia, à primeira vista, ser visto como um avanço, mas provocaria reações do sistema e seria de uma logística muitíssimo complicada, além de onerosa e inteiramente inútil para o caso brasileiro. A nossa expectativa, considerando o conhecimento de outros contextos e experiências pessoais, centrou-se na possibilidade de um exame, obrigatório para todos os aspirantes a estudos superiores, que tivesse alguma identidade com as grandes linhas do SAT – *Scholastic Aptitude Test*, desenvolvido e aprimorado no *Educational Testing Service* (Princeton, New Jersey, USA), e que, considerando-se as peculiaridades do nosso sistema educacional, tivesse diferentes

normas de interpretação, conforme veremos mais adiante.

A concretização da louvável ideia do ENEM – Exame Nacional do Ensino Médio – fez surgir alguns problemas que merecem discussão, a começar pelo seu próprio nome. Trata-se de um exame, circunstância que nos remete imediatamente à ideia de medida, que, eventualmente, pode ser usada numa avaliação, sem que isso, entretanto, signifique o começo necessário de toda e qualquer avaliação. Temos, também, um exame que não é obrigatório nos termos em que foi instituído; contudo, mecanismos de cautela foram criados para promover a sua aceitação e contornar resistências, que de fato vieram a ocorrer e ainda persistem. Alguns sistemas oficiais – *ça va sans dire* – assumiram o pagamento da taxa cobrada aos alunos e que era um dos motivos de oposição ao exame; posteriormente, os alunos carentes, certamente a grande maioria dos que frequentam o sistema público de ensino, ressalvados alguns bolsões da chamada classe média baixa, foram liberados dessa mesma taxa de inscrição. Ao conjunto de diferentes estímulos, para garantia da aceitação do exame, foi agregada a proposta, algo temerária, convenhamos, do uso dos seus resultados no acesso à seleção para o ensino superior, medida recebida com entusiasmo por algumas instituições e aceita com reserva por outras, inclusive oficiais, que passaram a admitir o resultado desse exame, mas, cautelosamente, fixaram alguma forma de ponderação, para evitar que os resultados do seu próprio processo seletivo fossem invalidados.

A aceitação do escore ENEM, para fins de acesso ao ensino superior, precisa ser cuidadosamente repensada, porque influencia no aumento do ponto de corte (e isso efetivamente ocorre, e vem ocorrendo, em vestibulares de primeira linha), sendo que, em alguns casos, esse acréscimo chega a ser acima de cinco pontos, tornando ainda mais elitista o processo de seleção para a Universidade e para algumas outras instituições de nível superior. É forçoso reconhecer que o uso do escore ENEM no vestibular acaba com o princípio da isonomia, porquanto dois estudantes, em igualdades de condições no processo seletivo, um, é favorecido, aquele que fez o ENEM, e o outro, ainda que com bons resultados, é preterido, simplesmente por não ter participado do ENEM.

O ENEM foi concebido para verificar competências e habilidades, segundo a formulação dos seus responsáveis, e pretende

avaliar cinco competências e vinte e uma habilidades, conforme reitera a sua literatura de divulgação. O assunto, evidentemente, não é pacífico, havendo contestações solidamente fundamentadas que apresentam dúvidas quanto ao conceito e à natureza dessas competências e habilidades. São dúvidas não necessariamente acadêmicas e que precisariam ser dirimidas, dada a sua complexidade. A situação se nos afigura bastante conflituosa, quando se observa que o próprio órgão responsável pela avaliação proclama, alto e em bom som, que o ENEM "não mede conteúdos, mas apenas competências e habilidades". Confessamos a nossa perplexidade e a forma dogmática da assertiva faz-nos lembrar a lição do mestre da Universidade de Chicago, Benjamin Bloom, injustamente esquecido entre nós, quando afirmava com bastante clareza que, ao avaliarmos um conteúdo, estamos, implicitamente, avaliando algo mais, as capacidades. Se considerarmos alguns exemplos, veremos que é impossível verificar a habilidade numérica de uma criança, sem constatar seus conteúdos de matemática; é impossível certificar a habilidade mecânica de um jovem, no conserto de um carro, por exemplo, sem considerar seus conteúdos de mecânica de automóvel; é inviável atestar a habilidade cirúrgica de um médico, sem considerar seus conteúdos de clínica médica, técnicas cirúrgicas e outros conteúdos mais ligados a uma determinada patologia.

Os princípios que baseiam o ENEM ficam comprometidos quando se examina o próprio instrumento utilizado, que parte de situações que demandam, liminarmente, conhecimentos de conteúdos, às vezes bastante complexos, e entendimento da sua verbalização, muitas vezes excessiva. Acreditamos que o ENEM poderia se tornar um instrumento eficiente de avaliação, e ser mais palatável para a sua clientela, assim como para a comunidade das instituições de nível superior, evitando contestações e confrontações, se ficasse restrito a apenas duas capacidades básicas, fundamentais na vida prática e indispensáveis em estudos superiores – a capacidade VERBAL e a capacidade NUMÉRICA, como veremos a seguir, na análise de três situações.

## TESTE DE APTIDÃO VERBAL E NUMÉRICA – A VERSÃO SAT

O *Scholastic Aptitude Test* – SAT é um instrumento desenvolvido a partir dos anos 20 e utilizado pelo *College Entrance Examination Board* – CEEB, nos Estados Unidos, para medir habilidades de raciocínio nas duas áreas anteriormente referidas: – verbal e numérica, conforme a apresentação de Donlon e Angoff (1971). Oferece escores separados para essas duas áreas e visa a verificar a competência dos estudantes que pretendem o ingresso em instituições de ensino superior. A função desse instrumento consiste em complementar informações, confirmando ou questionando, o desempenho em áreas de conteúdo, eliminando erros e inconsistências que possam ter ocorrido em avaliações anteriores restritas unicamente a conteúdos programáticos. É, reiteramos, um instrumento de habilidades básicas, cujos resultados vão integrar uma equação de regressão composta do SAT verbal, SAT numérico, escores do nível médio e outros elementos, não sendo usado apenas, e exclusivamente, o escore do SAT como um fator isolado, conforme crença de muitos. As pesquisas demonstraram que o SAT, que é uma medida padronizada em uma escala comum, possui alta validade preditiva dos melhores desempenhos nos *colleges* e nas universidades, acrescentando algo mais aos elementos de informação que integram a equação final usada para fins de seleção e classificação.

O SAT baseou-se na definição expressa por Ryans e Fredericksen (1951) e, sobretudo, na definição operacional de Cronbach (1960), com vistas a medir aspectos de habilidades desenvolvidas ao longo do tempo, fixando-se em habilidades verbais e numéricas, partindo do princípio de que as mesmas se constituíram no decurso da interação do estudante com o meio e, dessa forma, passaram a ser um equipamento relativamente independente da aprendizagem formal na escola. O conteúdo do SAT é balanceado a fim de compensar diferenças de interesses e de background dos vários segmentos da população. Ao longo dos anos, é necessário destacar, o SAT procurou introduzir outros elementos além do verbal e do numérico, mas nenhum deles demonstrou altas associações com desempenhos posteriores; desse modo, o SAT continuou identificado com a sua definição inicial centrada nos dois conjuntos de habilidades já referidas.

Ao longo dos anos, a parte verbal tem sido bastante diversificada, partindo de subsídios de diferentes áreas – social, política e científica – às quais são agregados elementos de outras áreas – literária, artística e filosófica –, enquanto a parte numérica do SAT procurou afastar-se de conteúdos curriculares, na medida do possível, concentrando-se em raciocínio lógico e na percepção de relações matemáticas. O SAT, ressalte-se, possui várias formas ou versões para aplicação em diversos momentos do ano, ao longo de anos sucessivos, e para fins de evitar problemas com a interpretação dos escores, são os mesmos padronizados em uma escala com média pré-fixada de 500 e desvio padrão igualmente preestabelecido de 100.

Vejamos a estrutura básica do SAT, conforme a descrição apresentada em Donlon e Angoff (1971), atentando, entretanto, para o fato de que, ao longo dos anos, o SAT vem sofrendo alterações bastante cautelosas e muito controladas, ao introduzir algumas poucas alterações no seu conteúdo e na apresentação de novos tipos de itens, considerando a complexa problemática do *equating* (tornar equivalentes resultados de diferentes versões do mesmo teste) e da estrutura fatorial do teste. A última alteração de que temos notícia foi a ocorrida no início da década de 90, conforme comunicação durante a reunião anual da *International Association for Assessment in Education*, realizada no Saint Patrick´s College, em Dublin (1992); assim sendo, a versão ora apresentada refere-se àquela que é analisada no relatório coordenado por William Angoff, inicialmente referido. Nesse formato, a parte verbal do SAT, composta de 90 itens, envolve antônimos, sentenças a completar, analogias e compreensão de leitura de textos. A parte numérica, com 60 itens, apresenta dois conjuntos de itens, sendo que um deles reflete questões habitualmente encontradas em testes de matemática e o outro usa itens sobre suficiência de dados. Os itens estão organizados em ordem de dificuldade crescente, igualmente padronizada pelo coeficiente Delta , a partir dos mais fáceis, em cada um dos blocos, e a dificuldade média de cada bloco é igual à dificuldade do teste no seu conjunto, o que é possível tendo em vista as cuidadosas estatísticas levantadas na fase de pré-testagem.

Os itens no SAT são de múltipla escolha, com cinco alternativas, e os folhetos de prova contêm alguns itens a mais (25), chamados de itens variantes, pois variam de aluno para aluno e de prova para prova, sendo que alguns desses itens va-

riantes destinam-se a obter informações necessárias à equalização das várias formas; outros, usados como se a aplicação fosse uma fase de pré-teste, serão incorporados mais tarde a futuras versões do SAT, e um terceiro conjunto de itens destina-se à realização de pesquisas. Esclareça-se, também, que os itens variantes não diferem dos demais itens operacionais. São itens paralelos, na medida do possível, com o objetivo de evitar a ocorrência de resultados enviesados (item bias) em relação a determinadas variáveis. A aplicação total do SAT é de três horas, sendo duas horas e meia para os itens operacionais e a restante meia hora para as questões variantes.

O SAT, ainda que seja um teste de aptidão, é, igualmente, um teste de desempenho (*achievement*), mas deste difere pelo fato de que é mínima a sua dependência em relação aos currículos tradicionais. Um aspecto a ressalvar na parte verbal relaciona-se aos itens de compreensão de textos, que são em número de sete e envolvem ciências biológicas, ciências físicas, humanidades, estudos sociais, havendo outros três itens que abrangem narração, síntese e argumentação. As questões estão distribuídas em três amplas categorias, que, por sua vez, são subdivididas em categorias mais restritas. Temos itens de COMPREENSÃO, abrangendo (1) compreensão da ideia principal e (2) compreensão de ideias secundárias; itens de RACIOCÍNIO LÓGICO, envolvendo (3) completar inferência pretendida, (4) o uso de generalização e (5) a avaliação da lógica da linguagem do texto; e, finalmente, itens relacionados a ASPECTOS EMOCIONAIS DA LINGUAGEM, (6) envolvendo a percepção do estilo e do tom do texto.

A dimensão conteúdo do subteste numérico do SAT abrange três categorias: aritmética-álgebra, geometria e "outros". A combinação de aritmética e álgebra resulta de que as regras básicas de combinação para ambas são as mesmas e, em muitos casos, os itens podem admitir uma solução por métodos aritméticos ou algébricos. A categoria geometria apresenta itens que demandam exclusivamente conhecimentos da geometria euclidiana dedutiva; por sua vez, a categoria "outros" inclui problemas que versam sobre lógica, topologia intuitiva, símbolos não usuais, operações e definições. Quanto às capacidades exigidas, os itens compreendem, habilidade computacional, julgamento numérico e estabelecimento de relações, além de outras mais classificadas como "miscelânea".

## OUTROS TESTES DE APTIDÃO VERBAL E NUMÉRICA – EXEMPLOS

Após as considerações sobre o SAT, veremos, em suas linhas gerais, a experiência do *Swedish Scholastic Test* (SweSAT), aplicado desde 1991 para fins de acesso às universidades na Suécia, abrangendo ampla gama de conteúdos e de níveis cognitivos, além de solicitar o desempenho em um subteste de Compreensão de Leitura em Inglês. A aplicação total do SweSAT, com 148 itens, é de quatro horas e o instrumento consta de seis subtestes, medindo habilidades verbais e não-verbais, uso de informações e conhecimentos de caráter geral, incluindo, ainda, compreensão de textos em inglês. A configuração geral do teste é a seguinte:

**(1)** o subteste PALAVRA – consta de 30 itens e mede a compreensão de palavras e conceitos;

**(2)** o subteste RACIOCÍNIO QUANTITATIVO – possui 20 itens e mede habilidades de raciocínio numérico na solução de problemas;

**(3)** o subteste COMPREENSÃO DE LEITURA - formado por 24 itens, mede a capacidade de compreensão de textos, sendo composto de quatro textos com seis itens cada um;

**(4)** o subteste DIAGRAMAS, TABELAS e MAPAS – engloba 20 itens e consiste em um conjunto de informações sobre um determinado assunto e a sua complexidade varia da interpretação de um gráfico à solução de problemas com dados de diferentes fontes;

**(5)** o subteste INFORMAÇÃO GERAL - compreende 30 itens, baseados em informações adquiridas ao longo dos anos de escolaridade, versando as mesmas sobre aspectos ligados ao trabalho, à educação, a problemas sociais, culturais e a atividades políticas;

**(6)** o subteste de COMPREENSÃO de LEITURA em INGLÊS, formado por 24 itens, possui uma formatação semelhante ao subteste de Compreensão de Leitura (3) e compreende de 8 a 10 textos de diferentes tamanhos.

O teste usa questões de múltipla escolha com quatro alternativas e suas funções básicas e características estão descritas por Wedman (1995), professor da Universidade de Ameä (Suécia), que também faz uma discussão sobre o seu desenvolvimento, uso e pesquisa em outro trabalho (1994)

Beller (1995), do *National Institute for Testing and Evaluation*, em Jerusalém, ao discutir os atuais dilemas e as soluções propostas para Israel, apresentou o esquema do *Psychometric Entrance Test* – PET (1990), construído com o objetivo de estimar sucesso em futuros estudos acadêmicos, que consta de três subitens:

**(1)** RACIOCÍNIO VERBAL – com 60 itens que, basicamente, procuram avaliar a habilidade de analisar e compreender material escrito de natureza complexa; a habilidade de pensar sistemática e logicamente, e a habilidade de distinguir o significado de palavras e conceitos. A parte verbal contém diferentes tipos de questões, como antônimos, analogias, completamento de sentenças, lógica e compreensão de leitura;

**(2)** RACIOCÍNIO QUANTITATIVO – possui 50 itens que procuram avaliar a habilidade de usar números e conceitos matemáticos na solução de problemas algébricos e equações, assim como em problemas geométricos. O subteste, além disso, verifica a capacidade de resolver problemas quantitativos e a de analisar informações apresentadas sob a forma de gráficos, tabelas e diagramas;

**(3)** a parte do subteste de INGLÊS avalia o domínio do inglês como segunda língua e os seus resultados integram o escore total do PET, servindo, também, para a organização de classes de recuperação para os que não têm um bom desempenho linguístico. O subteste consta de 54 itens, compreendendo sentenças para completar e reescrever, além de compreensão de textos.

Todos os itens do PET são de múltipla escolha e cada subteste é corrigido separadamente, numa escala padronizada com a média 100 e o desvio 20. O escore total do PET é a média ponderada dos escores nos três subtestes (40% Verbal; 40% Quantitativo e 20% Inglês), transformados numa escala padronizada com a média 500 e o desvio 100, variando os escores, assim como no SAT, de 200 a 800. O teste é apresentado nas seguintes línguas: – hebreu, árabe, espanhol, francês, inglês e russo, sendo os escores nas diferentes versões equalizados em relação aos resultados do teste em hebreu. Os candidatos que fizeram o teste em outra língua que não o hebreu devem fazer um teste de domínio nessa língua, por ser o hebreu a língua

oficial nas universidades. O artigo de Beller também analisa e esclarece três aspectos em relação ao PET – eficiência, viés e efeitos (pessoal, social e educacional).

## O ENEM – ALGUMAS QUESTÕES BÁSICAS

O instrumento usado no ENEM, tal como se apresenta no momento, carece de requisitos fundamentais, como mostra uma simples inspeção visual da distribuição dos itens, destacando-se, inicialmente, a validade de conteúdo. A essa deficiência, acrescenta-se outra, igualmente grave ou talvez mais grave ainda, por suas implicações, relacionada à validade de construto. O teste, medindo competências e habilidades, conforme sua literatura de divulgação, por sua própria natureza se baseia em construtos, mas, ao que nos consta, até a presente data não ofereceu evidências empíricas de que estaria efetivamente medindo aquelas variáveis que, supostamente, se propõe a medir. O teste, apesar dos esforços daqueles que participam da sua construção, salvo melhor juízo, não se fundamenta em dados empíricos sólidos, apoiados em pesquisas que não deixem dúvidas quanto à sua estrutura fatorial e a outros elementos oriundos de estudos psicométricos que evidenciem estar medindo aqueles atributos proclamados.

Existem numerosas metodologias já assinaladas há mais de trinta anos por Brown (1970) que poderiam ser utilizadas, inclusive a proposta por Campbell e Fiske (1959) que, comprovadamente, se adapta ao estudo dessa característica fundamental, já evidenciada há quase meio século por Cronbach e Meehl (1955), inicialmente, para os testes psicológicos, mas, depois, incorporada à teoria dos testes educacionais pelo próprio Cronbach (1971), no seu seminal ensaio sobre validação dos instrumentos de medida. Esse instrumento deve merecer aprofundados estudos psicométricos e discutidos os seus resultados, além de considerar suas múltiplas implicações educacionais, especialmente tendo em vista que há quem advogue o seu emprego em substituição ao atual processo de seleção para acesso a universidades e a outras instituições de ensino superior.

É preciso lembrar que, considerando a destinação do instrumento usado no ENEM, criado para medir competências e habilidades, deve o mesmo apoiar-se em uma teoria devidamen-

te comprovada do ponto de vista empírico. A verificação do seu funcionamento em relação a diferentes grupos é impositiva, sobretudo no caso nacional, que apresenta imensa diversidade social, econômica, cultural e educacional, oferecendo quadros bastante contrastantes. É sabido que os escores de um teste são influenciados por mudanças nos indivíduos e em decorrência de fatores ambientais, sendo que em nosso caso, numa mesma área geográfica, coexistem o 1º e o 3º Mundo, acentuando mais as gritantes disparidades regionais. Outro aspecto importante a verificar seria a constatação da não exigência de outras habilidades especiais, além das que supostamente estariam sendo medidas, para evitar turbulências que se podem refletir nas matrizes de correlações. Há exatos 20 anos, tentamos chamar a atenção da comunidade educacional para a relevância da validade de construto (Vianna, 1983), mas as coisas continuam como estavam em priscas eras. A inocência docimológica, assim como a inocência em educação, magistralmente analisada por Bloom (1976), ainda é uma realidade.

## AVALIAÇÃO E USO DE ESCALAS – O MITO DAS COMPARAÇÕES

A análise das grandes avaliações realizadas em território nacional, independentemente do nível administrativo que as promova, leva-nos a alguns problemas complexos e de difícil solução, como os relacionados às escalas empregadas, ao tipo de instrumentação usado e aos julgamentos comparativos que são emitidos sem maiores considerações sobre suas implicações e consequências decorrentes das repercussões no ambiente educacional e suas extrapolações na sociedade.

O uso de diferentes tipos de escalas não constitui problema, desde que seus referenciais apresentem pontos comuns que os tornem equivalentes, o que nem sempre ocorre. Assim, os grandes referenciais são quase sempre a média, o desvio padrão e o chamado escore "z", que expressa a relação da diferença entre o escore obtido e a média do grupo em termos de desvio padrão. Os escores passam a ter valores, teoricamente, entre menos 3,0 e mais 3,0, passando por 0,0, que corresponde à média. É evidente que, do ponto de vista técnico, essa escala oferece resultados sa-

tisfatórios para os especialistas, mas seria de difícil compreensão para a grande massa, sendo, então, transformada, acrescentando-se um fator multiplicativo pré-definido, o desvio padrão requerido, e um outro fator aditivo, igualmente pré-definido, a média desejada. Assim, a escala estaria linearmente padronizada, como no caso de 10z + 50, em que os escores variariam de 20 a 80, ou um escore 100z + 500, com valores variando de 200 a 800, sendo a média no primeiro caso igual a 50 e no segundo a 500, como acontece no SAT e em outros testes cujos escores são padronizados, inclusive em avaliações internacionais em larga escala.

Apresentamos uma visão simplificada do escore padronizado para encaminharmos a nossa discussão e chegarmos a um ponto crítico em relação às avaliações do MEC com as suas escalas de proficiência, com níveis que vão de 125 a 400, com intervalos de 25 pontos. As informações nem sempre claras dos relatórios não nos permitem entrar em maiores detalhes sobre o processo de padronização das escalas. Uma pergunta, associada a essas escalas de proficiência, nos veio à mente: – será razoável colocar centenas de milhares de sujeitos em uma única escala (ainda que com base na chamada Teoria da Resposta ao Item (TRI) isso seja estatisticamente possível), ignorando completamente a diversidade social, econômica, cultural e educacional dessa população e as distorções que influenciam a caracterização dos vários índices de desenvolvimento humano? Não seria razoável, considerando as variáveis apontadas, construir normas diferenciadas por região, levando em conta a diversidade das características individuais? Talvez, a título de sugestão, fosse o caso de termos uma norma para cada uma das regiões geoeconômicas, fazendo-se alguns ajustamentos em certos casos, como no Sudeste e no Sul. Pensamos que se poderia ter uma visão menos distorcida da realidade brasileira, desde que as escalas tivessem os mesmos referenciais, relacionados às médias e aos desvios padrão de cada área regional, criando-se, desse modo, uma geografia da educação, a exemplo do que é feito na França, inclusive com a incorporação dos valores agregados que ressaltariam o papel da educação, especialmente nas regiões em que as desigualdades sociais são mais acentuadas.

Antes de voltarmos ao problema das comparações, ao mito das comparações, para usarmos a expressão de Nuttall (1995), mostraremos a nossa dúvida sobre como classificar o tipo de

avaliação a que se propõem o SAEB e o ENEM. A dúvida que nos assalta é se seria uma avaliação referenciada a norma ou referenciada a critério. O problema decorre do fato de que, pelo esquema de planejamento, por sua estrutura final, pelos processos de correção, entre outros elementos, tudo nos leva a crer que se trataria de um instrumento referenciado a norma, ao desempenho do grupo, refletido em diferentes tipos de estatísticas; contudo, quando observamos as escalas de proficiências e vemos as diferentes habilidades referenciadas a diferentes níveis específicos de desempenho (critérios), ficamos na dúvida – norma ou critério? –, dúvida, aliás, que não é exclusivamente nossa, tendo sido inclusive objeto de consideração no Grupo de Trabalho sobre Padrões e Avaliação do PREAL (Programa de Promoção da Reforma Educativa na América Latina e no Caribe), no fórum de discussão sobre As políticas de avaliação do desempenho da aprendizagem nos sistemas educativos da América Latina (2003).

Voltando ao problema das comparações, perguntamo-nos – qual o seu significado, qual é, efetivamente, o seu objetivo? Quando ouvimos alguém dizer, por exemplo, que o desempenho de um aluno da 3ª série do ensino médio no vale do Gurupi corresponde ao desempenho de um aluno de 8ª série do ensino fundamental do vale do Itajaí, acreditamos que a comparação se faça simplesmente pelo hábito de comparar, pois dessa comparação nada efetivamente resulta, salvo maliciosos comentários de alguns segmentos da mídia, tendo em vista suas implicações. Como comparar um indivíduo que vive numa zona de economia extrativista, numa área de índices sociais comprometidos, com um outro sujeito de uma região com economia bem próxima da existente no primeiro mundo e com altos índices sociais positivos?

Além de aspectos sociais e econômicos, precisamos atentar para a diversidade das características dos sistemas educacionais em diferentes regiões, a natureza dos currículos, a formação e experiência do corpo docente. Diante desse quadro, podemos fazer comparações e imaginar que os indivíduos poderiam ter os mesmos conhecimentos e as mesmas capacidades? É bom lembrar, fazendo referência novamente a Nuttall (1995), que a comparação entre padrões não significa, necessariamente, identidade de desempenhos. O ato de comparar tem muito pouco de certeza, não se constitui em um procedimento de rigorosa análise estatística. A comparação resulta de um julgamento humano, sujeito, dessa

forma, à falibilidade, considerando, também, que o conceito de comparar é extremamente vago. Apesar de tudo, comparar tornou-se um ato obsessivo na prática de algumas avaliações – são comparados sistemas, desempenhos por disciplina, comparam-se disciplinas ao longo dos anos e o mesmo procedimento é adotado em relação a diferentes programas –, chegando a um lamentável e absurdo exercício, por ignorar o fato de que qualquer avaliação de um ser humano é feita por um outro ser humano e os escores resultantes nunca se revestem de uma precisão absoluta, que demandaria instrumentos perfeitos isentos de erros de medida, o que é impossível na prática, mesmo que utilizadas tecnologias de ponta e processos estatísticos sofisticados.

## EXAME NACIONAL DE CURSOS – ENC – UMA GRANDE CONTROVÉRSIA

Chegamos, nesta fase da presente reflexão, a um terceiro momento da discussão sobre a avaliação da educação brasileira – o Exame Nacional de Cursos – ENC – para as instituições de Ensino Superior, públicas e privadas, compreendendo Universidades, Centros Universitários, Faculdades Integradas e instituições isoladas de ensino de terceiro grau. O ENC foi chamado pela massa estudantil de Provão, denominação esta incorporada pelos órgãos oficiais da educação, que a adotaram inclusive como título de uma revista de divulgação dos seus pressupostos e objetivos. O novo Exame Nacional de Cursos, que vigora a partir de 1996, sendo obrigatório para todos os alunos formandos, por força de instrumento aprovado pelo Congresso Nacional, nasceu sob o signo da contestação de alguns segmentos, inclusive professores e alunos, mas foi, entretanto, inteiramente aceito pela sociedade, que passou a utilizar seus resultados para fins de escolher cursos nas instituições mais bem situadas na classificação final, baseada parcialmente no desempenho dos alunos em instrumentos de verificação do rendimento acadêmico. Houve nisso um grande equívoco, pois o critério de avaliação das instituições não se restringe apenas a provas, inclui, também, a avaliação do corpo docente, a do projeto pedagógico e a da infraestrutura institucional, que, juntamente com o Exame Nacional de Cursos, resultam na Avaliação das Condições de

Ensino. O chamado Provão é apenas uma das dimensões de um processo mais amplo (e bastante controverso, como veremos).

A avaliação do ensino superior constitui, sem sombra de dúvida, uma necessidade. O crescimento do atual Ensino Básico, desde os anos 60, e a nova configuração da rede de ensino, inclusive com o justo aumento dos anos de escolaridade obrigatória, entre outros elementos, contribuíram para o surgimento de pressões sobre o nível de escolaridade subsequente, promovendo, assim, a eclosão de numerosas faculdades e a abertura de novos cursos em diferentes instituições, sobretudo privadas, em um ritmo inteiramente descontrolado. Ao aumento quantitativo corresponderam dúvidas quanto à qualidade do ensino, à eficiência do corpo docente e à devida adequação das condições institucionais, que justificaram a ação governamental, ainda que tardia.

A criação do ENC teve de imediato grande repercussão no ensino privado, que se viu diante de uma situação inédita no quadro educacional brasileiro, e gerou, igualmente, reações no ensino público, especialmente tendo em vista a argumentação, nem sempre defensável, da autonomia universitária, que estaria sendo violada. Alguns problemas não foram realmente definidos com a devida adequação, destacando-se, entre outros, a mal dimensionada obrigatoriedade do Exame para todos os alunos formandos sem a fixação de uma nota de corte, que refletisse um nível mínimo de competência desejável. A falta de um escore mínimo fez com que prevalecesse simplesmente a presença do aluno, independentemente do seu desempenho. Isso, traduzido em termos de ação, significou que muitos estudantes contrários ao exame, por motivos vários, inclusive ideológicos, se limitassem a assinar o documento comprovante da sua presença – a folha de respostas da prova – e ignorassem o conteúdo curricular exigido, entregando a prova em branco ou nela expressando protestos, e garantindo, dessa forma, a expedição do diploma, tendo em vista o atendimento do ritual legal.

A diversidade dos numerosos cursos a serem avaliados levou o MEC a constituir comissões que definissem para cada prova as várias áreas objeto do Exame e estabelecessem uma certa "filosofia" para cada uma das avaliações, segundo a proposta oficial de verificar os conhecimentos fundamentais necessários aos formandos de cada curso. Vimos, desse modo, que certas definições envolveram elementos dos cursos básicos ministrados nos pri-

meiros momentos da sequência formativa, omitindo ou deixando de considerar outros aspectos objeto de estudos nas últimas séries da formação acadêmica. Além do mais, seria preciso que o MEC levasse em consideração o fato de que similaridades curriculares nem sempre traduzem identidades e cursos com a mesma designação podem ter estruturas inteiramente diferenciadas; desse modo, na prática, os "*syllabus*" – se assim podemos chamar –, que foram divulgados pelo MEC, e são dados a conhecer todos os anos, na época do Exame, passaram a ser programas de "ensino" em muitas instituições, mais preocupadas com o que seria a avaliação institucional do que com a formação geral, científica e profissional do seu alunado. Além do mais, algumas instituições, considerando as repercussões do desempenho dos alunos no seu "*marketing*" promocional, desenvolveram imaginosas estratégias de "ensino" com vistas ao preparo para o ENC ou, mais especificamente, para o hoje célebre "Provão", configurando-se nova modalidade de "cursinho" preparatório.

Outras comissões, integradas por membros de diferentes instituições, necessitam ser organizadas ao longo do processo de desenvolvimento do ENC. Assim, definidos os conteúdos, constituem-se grupos para a elaboração dos instrumentos, ressaltando-se que estes novos grupos são diferentes dos que definiram a "filosofia" e desenvolveram o que chamamos de "*syllabus*". Apresentam-se muitas vezes situações conflitivas, pois os que devem elaborar o material do Exame nem sempre têm as mesmas percepções teóricas dos que integraram a primeira comissão, dificultando, desse modo, a operacionalização do Exame. É bem possível, a título de uma exemplificação inteiramente hipotética, mas não absurda, que um grupo junguiano deva implementar uma programação de sabor skinneriano ou vice-versa; ou que um programa de física orientado no sentido eminentemente experimental deva ser trabalhado por um outro grupo extremamente matematizado ou vice-versa; ou que um programa de biologia inspirado na química molecular deva ser operacionalizado por um grupo mais chegado a uma orientação tradicionalista ou vice-versa. Essas são algumas hipóteses levantadas para configurar situações que podem ser consideradas impossíveis, mas que ocorrem na prática do dia a dia, em que divergências conceituais, filosóficas e de tratamento dos vários assuntos existem, sem dúvida, dificultando

ou mesmo impossibilitando o trabalho dos responsáveis pela definição operacional dos vários conteúdos a examinar.

Ainda com relação a conflitos entre o grupo que idealiza um esquema e o que constrói os instrumentos, podemos imaginar o seguinte: – suponhamos que o grupo idealizador, imbuído da ideia traduzida no binômio ensino/pesquisa, aliás discutida recentemente com bastante equilíbrio por Moura e Castro (Veja, 22.12.02), resolva exigir a elaboração de um "projeto de pesquisa", numa situação de exame como o que ora é analisado. Como operacionalizar esse mito educacional denominado "ensino/pesquisa" numa situação artificial de "stress" que envolve milhares de pessoas que trabalham sem fontes de consulta e de referência dentro de um período de tempo bastante restrito? A situação proposta não é tão estranha quanto pode parecer a um primeiro exame. A solução desse conflito poderia ser superada pela atuação conjunta das duas comissões – a que teoriza e a que implementa –, que se proporiam a elaborar um programa que traduzisse um certo consenso, admitindo-se que seja possível um consenso em questões educacionais.

Antes de referirmo-nos a uma terceira comissão participante do ENC, queremos analisar aspectos ligados a pequenas comissões, integradas por funcionários do MEC e/ou por pessoas da confiança do Ministério, que fazem a revisão formal das questões ou dos itens, depois de pronto o instrumento e revisto pela própria comissão elaboradora e por um revisor especialista na área. A comissão do MEC procura seguir de uma forma bastante ortodoxa princípios definidos ao longo dos tempos por psicometristas e algumas instituições especializadas, como o *Educational Testing Service* (Princeton, New Jersey), e disseminados por pessoas direta ou indiretamente ligadas a centros de pesquisa e avaliação, quase sempre norte-americanos. O excesso de formalismo, queremos acentuar, nem sempre traz grandes contribuições, mas quase sempre constitui fator de perturbação, devendo prevalecer o bom senso no uso de pequenas regras, que se podem transformar em verdadeiros preciosismos, quando usadas sem as devidas cautelas.

Definidos os objetivos da avaliação, estabelecidos os parâmetros para a elaboração dos instrumentos, discutidas, revistas e aplicadas as provas com a posterior divulgação dos resultados, inicia-se, na dinâmica do ENC, a atuação de uma nova comissão

com elementos que não participaram das várias fases anteriores, com o objetivo de, em princípio, fazer uma análise crítica dos instrumentos elaborados. É sabido que não existem instrumentos perfeitos, especialmente no caso presente, pois medem elementos não tangíveis que englobam aspectos cognitivos e diferentes capacidades relacionadas ao construto que, supostamente, está sendo mensurado. Toda e qualquer discussão na área é sempre proveitosa, dependendo dos seus termos e, no caso presente, as considerações devem basear-se nas matrizes compostas por diferentes elementos estatísticos possíveis de coletar sobre o desempenho dos que responderam às questões. Isso não significa, ressaltemos, que não haja um certo subjetivismo sempre que são expressos juízos de valor relacionados a assuntos e à maneira como foram abordados nas várias questões; entretanto, esse subjetivismo não pode resultar de posicionamentos ideológicos, idiossincrasias pessoais e nem decorrer de antagonismos acadêmicos. O que se observa, no entanto, é que essas discussões possuem um tom eminentemente impressionista – eu acho; eu penso; eu acredito; eu julgo – sem qualquer tipo de fundamentação empírica ou teórica; por outro lado, as críticas não incidem sobre o instrumento como tal, sua estrutura, seus possíveis e até mesmo compreensíveis defeitos, mas resultam de um posicionamento muitas vezes contrários à filosofia, à prática do Exame Nacional de Cursos e à sua razão de ser, refletindo, por outro lado, um certo antagonismo a toda a política educacional que fundamentou a decisão de instituir um amplo programa de avaliação de todo o sistema educacional do país. A análise supostamente crítica reflete com bastante frequência um certo sabor xenófobo, digamos, ao considerar o instrumento com um viés regional, considerando a prova como identificada com certas instituições, mas negando-lhe valor em relação a outras.

## O EXAME NACIONAL DE CURSOS E O USO DA CURVA NORMAL

A presente consideração do ENC nos leva de um ponto crítico a outro, às vezes bem mais crítico que os anteriores, como é o caso do que ora passamos a considerar: – a apresentação inicialmente feita dos resultados do ENC expressos por concei-

tos associados a porcentagens fixas de tal forma que sempre teríamos, independentemente da distribuição dos escores, os conceitos A, B, C, D e E, com o mesmo número percentual de sujeitos em A e E, o mesmo número também percentual de elementos em B e D, e a maior concentração de estudantes na faixa do conceito C, refletindo, assim, a crença mítica na curva normal gaussiana, como se esta efetivamente traduzisse a distribuição das diferenças individuais. O uso da ideia da curva normal de Gauss, que nada mais é do que a expressão de uma determinada função matemática associada a grandes números e a fenômenos probabilísticos, foi uma tragédia de grandes proporções e da qual parte significativa do mundo da educação ainda não conseguiu se refazer. Diferentes tipos de curvas podem ser obtidos, dependendo da construção dos instrumentos e do grau de dificuldade dos itens (CRONBACH; WARRINGTON, 1952) e críticas à curva normal para explicar variáveis educacionais (e psicológicas) foram devidamente dimensionadas por Cronbach (1971 e 1977) e por Bloom, Hastings e Madaus (1971), sendo que estes três últimos colocaram a questão nos seguintes termos:

> Como educadores usamos a curva normal na atribuição de notas aos estudantes há tanto tempo que passamos a nela acreditar. Medidas do desempenho são planejadas para detectar diferenças entre nossos alunos – ainda que as diferenças sejam sem importância em termos de conteúdos. Então, distribuímos nossas notas segundo a curva normal. Em qualquer grupo de estudantes esperamos que uma pequena porcentagem receba A. Ficamos surpresos quando o número de alunos difere muito de cerca de 10 por cento. Estamos também preparados para que igual proporção de alunos fracasse. Muito frequentemente esse fracasso é determinado pela posição dos estudantes no seu grupo e não pela incapacidade de perceber as ideias fundamentais do curso. Assim, acostumamo-nos a classificar os alunos em cerca de cinco níveis de desempenho e a atribuir graus de uma maneira relativa. Não importa que os fracassados de um ano tenham o desempenho aproximado do nível daqueles que obtiveram conceito C no outro ano. Nem importa que os estudantes de nível A de uma escola tenham um desempenho igual ao dos estudantes que receberam F em outra escola. (p. 44-45)

É evidente que, como as distribuições dos resultados não apresentam uma normalidade perfeita e nem mesmo aproximada, mas, ao contrário, uma assimetria acentuada para a direita, positiva, com a maior concentração de escores baixos, o fato de um curso ter conceito A ou B não significa, necessariamente, pelo critério adotado, a excelência dos resultados; ao contrário, a maioria dos resultados A poderia situar-se abaixo da média teórica de 50, numa escala de 0 a 100. Tendo em vista, portanto, a bizarra mas não rara situação que se configurava com proporções pré-definidas para cada faixa conceitual, o MEC alterou seus critérios, tomando a média de cada curso em função da média e do desvio da totalidade dos cursos para estabelecer seus conceitos, conforme se pode ver no texto adiante reproduzido:

> O critério parte da média aritmética das notas dos estudantes que fazem o exame e considera a média geral da área e o desvio padrão, que mede a dispersão das notas em torno da média. Com isso, o conceito A é atribuído a todos os cursos que obtêm notas acima de 1.0 desvio padrão da média geral. O conceito B, aos que têm entre 0.5 e 1.0 desvio padrão acima da média geral. O conceito C vai para as faculdades que tiraram entrem 0.5 desvio padrão abaixo e 0.5 desvio padrão acima da média geral. Por fim, os cursos que ficam com os conceitos D e E têm notas entre 0.5 e 1.0 desvio padrão abaixo da média geral (D) e notas abaixo de 1.0 desvio padrão da média geral.

Verifica-se, dessa forma, que pode haver casos em que não existirão conceitos A e B, mas apenas conceitos C, D ou E, o que representou um certo avanço, ainda que não muito significativo, e persistiram ainda insatisfações, inclusive com recursos ao Poder Judiciário para impedimento da divulgação dos desempenhos dos cursos, o que se configura, mais uma vez, uma situação extremamente surpreendente, sobretudo tendo em vista o atendimento de liminar ao pedido. Lamentavelmente, no fundo, continuou a subsistir a ideia (e a fervorosa crença) de que a célebre curva normal traduz a distribuição de variáveis ligadas ao desempenho dos seres humanos.

## O PAPEL DO ESTADO EM AVALIAÇÕES – POSSÍVEIS ALTERNATIVAS

O Estado como avaliador sofre bastante restrições, mas não restam dúvidas de que uma avaliação, para fins de atestar a competência ao término de um curso, é algo que se impõe, inclusive com o apoio generalizado da sociedade. Acreditamos que existam soluções satisfatórias, vivenciadas em outros países e, em algumas situações, no próprio Brasil: – a avaliação por órgãos de classe, que podem exigir a comprovação da eficiência de uma pessoa para o exercício de determinada profissão, credenciando-a, após resultados satisfatórios, para a atuação em determinada área de conhecimento profissional selecionada para atuação na sociedade. A Ordem dos Advogados do Brasil, por exemplo, no caso da seção de São Paulo, realiza, anualmente, um exame pós-curso, a que todos os formandos em direito estão sujeitos, fato este que lhe permite, inclusive, identificar os cursos mais eficientes e os de menor sucesso, evitando, assim, que sejam lançados no mercado de trabalho milhares de futuros profissionais sem as requeridas qualificações. A excelência dessa medida estaria ligada à sua validade local, por Estado, ou seja, alguém, mesmo aprovado em um estado, ao se transferir para outro, seria obrigado a submeter-se a novo exame junto ao órgão local, evitando-se tentativas de burla a dispositivos que venham a regular a matéria. Outros exemplos podem ser citados na área médica. Alguns órgãos corporativos, como a Sociedade Brasileira de Pediatria e a Sociedade Brasileira de Ortopedia e Traumatologia realizam exames anuais, por intermédio dos quais atestam a capacidade de especialistas em suas respectivas áreas, e muitos hospitais já começam a exigir essa titulação para o exercício profissional em seu quadro médico.

Acreditamos que o exame de competência profissional e, implicitamente, da competência dos cursos superiores poderia ser realizado com bastante eficiência pelos órgãos corporativos regionais das diferentes profissões, sob o controle do seu respectivo órgão central. A aplicação de exames de competência deveria ser de responsabilidade dos órgãos corporativos regionais, que, inclusive, poderiam atuar em associação com outras instituições de direito privado especializadas em avaliação de recursos humanos qualificados, para fins de elaboração dos instrumentos, quando

fosse o caso. A certificação de concluintes de cursos de licenciatura ligados ao magistério poderia ser feita pelas Secretarias de Estado da Educação, com validade restrita aos seus respectivos estados.

## AUTOAVALIAÇÃO E AVALIAÇÃO EXTERNA – SEU SIGNIFICADO

Pensamos que essas e outras sugestões tenham praticabilidade e possam vencer ou atenuar as resistências ora oferecidas. Ao MEC e às Secretarias de Estado da Educação caberiam a importante e significativa missão de controlar os resultados das avaliações e aplicar as possíveis punições às instituições que não atingissem os parâmetros desejados. O assunto é polêmico, temos plena consciência, assim como quase tudo em educação é igualmente polêmico ou objeto de polêmicas. É preciso lembrar, além dos problemas anteriormente apontados, os atuais custos elevados do ENC e tememos que, em futuro bem próximo, seja o mesmo inviabilizado do ponto de vista financeiro. O assunto deve ser discutido pela sociedade, inclusive considerando outras alternativas além das que foram anteriormente propostas, a fim de alterar a atual situação, considerando que as próprias instituições de terceiro grau precisam de informações consistentes que lhes permitam aprimorar os seus procedimentos e atender a suas necessidades. A sociedade, sem dúvida, necessita, igualmente, de informações válidas e consistentes para julgar de forma criteriosa as instituições que, de um modo ou de outro, são suas subsidiadas.

A avaliação institucional de Universidades, Centros Universitários, Faculdades Integradas e de todas as modalidades de Instituições de Ensino Superior – IES que possam existir no sistema educacional brasileiro, salvo melhor juízo, deve basear-se, necessariamente, na AUTOAVALIAÇÃO e em AVALIAÇÕES EXTERNAS por iniciativa das próprias instituições, a exemplo do que já ocorre em algumas universidades que tiveram um papel pioneiro nessa iniciativa, como a Universidade Nacional de Brasília – UnB – e em outras instituições mais, que, sendo subordinadas a Conselhos Estaduais, como as universidades estaduais do Estado de São Paulo e os Centros Universitários de Santo André e São Caetano, no mesmo estado, já promovem suas autoavaliações.

É preciso resgatar a promissora experiência do Programa de Avaliação Institucional das Universidades Brasileiras – PAIUB, que, lamentavelmente, não foi levada adiante.

A autoavaliação e as possíveis avaliações externas, quando estas últimas se fizerem necessárias, a juízo das instituições, deveriam ser complementadas com avaliações eminentemente qualitativas dos programas de pesquisas pelas agências financiadoras, como, por exemplo, o CNPq e a FAPESP, e, finalmente, a avaliação também qualitativa, mas incluindo elementos quantitativos, dos cursos de pós-graduação pela CAPES, o que já vem ocorrendo. As autoavaliações, realizadas em intervalos a serem fixados, cinco anos, suponhamos, juntamente com possíveis avaliações externas para fins específicos, e mais os trabalhos de auditoria no campo da pesquisa e da pós-graduação, forneceriam, sem dúvida, elementos preciosos para o MEC exercer sua função principal de agência controladora da qualidade do ensino superior, podendo, inclusive, através de procedimentos legais apropriados, isentar alguns cursos de graduação de novos exames, a partir dos dados informativos oriundos dos órgãos corporativos responsáveis pelos exames de fim de curso, como a OAB, CFM, CREAs e outros conselhos mais, que tivessem comprovado de forma indiscutível a eficiência ao longo de quatro anos seguidos, suponhamos.

As presentes considerações, acompanhadas de algumas sugestões, que julgamos realistas face o atual quadro, visam a propor uma nova formatação às pioneiras avaliações em larga escala promovidas nos anos 90 pelo MEC e implementadas com grande eficiência pelo Instituto Nacional de Estudos e Pesquisas Educacionais – INEP. Queremos, ao finalizar, reiterar o significado da avaliação no processo educacional, como o fez Kellaghan (2001), e destacar sua importância no sentido de (1) elevar os padrões de ensino muitas vezes bastante comprometidos em algumas instituições; (2) ajustar os processos de ensino à aprendizagem com o uso de metodologias adequadas e que devem ser de domínio dos professores, o que nem sempre ocorre; (3) contribuir para a formação de cidadãos que possam desafiar a complexidade de uma sociedade tecnológica; e, ainda, (4) proporcionar aos responsáveis pela tomada de decisões educacionais o *feedback* necessário para que prevaleça o bom senso que, na prática, conduz ao acerto das ações.

## REFERÊNCIAS BIBLIOGRÁFICAS


BELLER, Michal. Admission to higher education: current dilemmas and proposed solution. In: KELLAGHAN, Thomas (Ed.). *Admission to higher education*: issues and practice. Dublin: Educational Research Centre; New Jersey: International Association for Educational Assessment, 1995.

BLOOM, Benjamin S. Inocência em educação. *Cadernos de Pesquisa*, São Paulo, n. 16, p. 63-71, mar. 1976.

BLOOM, Benjamin S.; HASTINGS, J. Thomas; MADAUS, George F. *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill Book, 1971.

BROWN, Frederick G. *Principles of educational and psychological testing*. Illinois: The Dryden, 1970.

CAMPBELL, Donald T.; FISKE, Donald W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, n. 59, 1959.

CRONBACH, Lee J. *Essentials of psychological testing*. 2th ed. New York: Harper and Row, 1960.

_____. Test validation. In: THORNDIKE, Robert L. *Educational measurement*. 2th ed. Washington, D.C: American Council on Education, 1971.

_____. *Essentials of psychological testing*. 3th ed. New York: Harper and Row, 1977.

CRONBACH, Lee J.; MEEHL, Paul F. Construct validity in psychological tests. *Psychological Bulletin*, n. 52, 1955.

CRONBACH, Lee. J.; WARRINGTON, Willard G. Efficiency of multiples: choice tests as function of spread of items difficulties. *Psychometrika*, n. 17, 1952.

DONLON, Thomas F.; ANGOFF, William H. The Scholastic aptitude test. In: ANGOFF, W.H. (Ed.). *The College board admissions testing program*: a technical report on research and development activities relating to the SAT and achievement tests. New York: College Entrance Examination Board, 1971.

KELLAGHAN, Thomas. The use of assessment in educational reform. In: ANNUAL CONFERENCE OF THE INTERNATIONAL ASSOCIATION FOR EDUCATIONAL ASSESSMENT, 27., 2001, Rio de Janeiro, *Paper*... Rio de Janeiro: IAEA, 2001.

NUTTALL, Desmond. The Myth of comparability. In: MURPHY, Roger; BROADFOOT, Patricia. *A Tribute to Desmond Nuttall*. London: The Falmer, 1995.

RYANS, D. G.; FREDERICKSEN, N. Performance tests of educational achievement. In: LINDQUIST, E. F. (Ed.). *Educational measurement*. Washington, DC: American Council on Education, 1951.

VIANNA, Heraldo M. Validade de construto em testes educacionais. *Educação e Seleção*, São Paulo, n. 8, p. 35-44, jul./dez. 1983.

WEDMAN, Ingeman. Selection to higher education in Sweden. In: KELLAGHAN, Thomas (Ed.). *Admission to higher education*: issues and practice. Dublin: Educational Research Centre; New Jersey: International Association for Educational Assessment, 1995.

# AVALIAÇÃO EDUCACIONAL: VIVÊNCIA E REFLEXÃO[1]

## INTRODUÇÃO

Ao longo de vários momentos, esforçamo-nos para definir a posição metodológica de alguns responsáveis pela criação da nossa concepção de avaliação, focalizando, especialmente, Tyler, Cronbach, Scriven, Stufflebeam e Stake. Procuramos destacar aqueles que se preocuparam com a lógica do processo (Scriven) e visaram a apresentar soluções para problemas práticos (Cronbach), criando, assim, uma sistemática para lidar com os problemas enfrentados pelos avaliadores. Chegamos, agora, ao momento de considerar algumas outras questões, refletindo o nosso posicionamento. Ao fazê-lo, partimos da lembrança das 95 teses apresentadas por Cronbach[2], esclarecendo que a presente reflexão representa um esforço visando a mostrar a compreensão que temos dos problemas considerados e vivenciados por nós no período de 1962 a 1995. É uma tentativa, portanto, de registrar o nosso pensamento sobre avaliação educacional de forma sistematizada, à luz de experiências pessoais, refletindo algumas das influências que sofremos ao longo daquele período e da prática da avaliação.

**1** Artigo publicado em *Estudos em Avaliação Educacional*, n. 18, p. 69-109, jul./dez. 1998.

**2** Cronbach, L I. et al. (1980) – à maneira de Martinho Lutero, com suas Teses (1517) que provocaram o Concílio Tridentino – divulgaram um conjunto de proposições advogando uma nova visão da avaliação, especialmente de programas sociais, criando uma nova consciência em re!ação à avaliação.

## AVALIAÇÃO E AUTOCONHECIMENTO

A educação não se limita a transmitir conhecimento elaborado e sistematizado ao longo do tempo. Esse conhecimento é indispensável à sobrevivência do indivíduo e sua necessidade se torna cada vez mais crítica à medida que a nossa cultura sofre um maior aprofundamento tecnológico, exigindo o domínio de técnicas necessárias a atividades bastante simples. A educação, entretanto, não fica restrita apenas a esse aspecto, pois assim não teria sentido o seu proceder. Para completar e dar significado ao seu proceder, impõe que, ao lado do conhecimento humanístico e tecnológico, sejam incorporados valores e tradições definidos culturalmente. Ora, diante desse quadro, a avaliação não é um processo que leva apenas ao levantamento de dados, à identificação de problemas para que os executivos da educação lhes deem as melhores soluções. A avaliação, vista sob o enfoque que pretendemos, é bem mais do que apenas isso; é uma forma de autoconhecimento da própria sociedade, que procura conhecer a si mesma através da identificação do que prevalece em uma de suas principais instituições – a escola – , que é responsável por sua continuidade. O fracasso refletido através de uma avaliação representa, na realidade, o próprio fracasso da sociedade em concretizar seus objetivos, ficando, desse modo, detentora de uma forma de conhecimento que poderá utilizar, mas que pouco contribuirá para a sua sobrevivência – uma escola falida reflete, através da avaliação, uma sociedade igualmente fracassada nos seus objetivos.

## AVALIAÇÃO E PLANOS DE AÇÃO

Algumas sociedades possuem verdadeira obsessão relativamente à avaliação – caso específico da norte-americana – ; outras, a usam com moderação, preocupadas muitas vezes com suas implicações sociológicas – como parece ocorrer na Inglaterra – , tornando o seu uso bastante cauteloso e um terceiro grupo, geralmente integrado por países do Terceiro Mundo, que sofrem a influência de agências internacionais de financiamento, as quais usam a avaliação como forma de controle dos seus subsídios financeiros e tentam influenciar políticas públicas. A situação nestes países é de plena euforia avaliativa, especialmente

diante da possibilidade de virem a conhecer seus sistemas de ensino e criarem novas condições para a superação de seus diversos problemas, geralmente graves, diríamos mesmo gravíssimos. Enquanto os primeiros elementos fazem investimentos de milhões em diversas moedas fortes; os últimos financiam seus empréstimos, muitas vezes a juros extremamente onerosos. Avaliações são feitas, às vezes, com o emprego de técnicas bastante sofisticadas e que nem sempre estão ao alcance dos que seriam os maiores interessados, pelo menos teoricamente, – os professores. Uma pergunta surge-nos: – o que ocorre com essas avaliações, particularmente as realizadas no mundo em desenvolvimento, com especial referência a alguns países do Terceiro Mundo? As avaliações são discutidas em círculos restritos, profissionalmente interessados, às vezes com a participação de elementos burocráticos, mas com repercussões em segmentos limitados. A partir das avaliações, e ao contrário do que seria desejável e recomendável, e mesmo impositivo, não são geradas novas discussões que levam a planos para uma ação efetiva. Às vezes ocorrem ações setoriais – revisão de currículos e os inevitáveis e frustrantes treinamentos de professores, via cursos que não correspondem às suas expectativas e necessidades reais. Na verdade, não são feitos planos alternativos de ação, que conduzam à mudança e à transformação da escola em termos positivos. Avaliação, sem o complemento de planos de ação, tende a ser uma atividade burocrática, que aos poucos perde o seu significado e passa a ter efeitos nulos e gera reações adversas, que se propagam em cadeia dentro do sistema. A avaliação passa a ter um efeito negativo, contrário ao que seria desejável.

## AVALIAÇÃO E TRANSFORMAÇÃO

A avaliação é às vezes vista de forma bizarra, como um processo revolucionário que promoverá mudanças imediatas em todo o sistema, abrangendo todos os sujeitos – administradores, técnicos, professores e alunos, com a alteração imediata de programas, projetos e materiais. Assim como na natureza nada se faz aos saltos, mas resulta de um processo às vezes razoavelmente longo, o mesmo ocorre em educação, especialmente em relação à avaliação e aos seus propósitos. A escola tende a ser conserva-

dora, visando, naturalmente, a sua sobrevivência, inclusive com os seus males; entretanto, a avaliação não está comprometida com esse conservadorismo, que gera um processo de estagnação e autodestruição. A avaliação admite um certo gradualismo e que as transformações venham a se processar progressiva e refletidamente, para evitar fracassos, alterações injustificáveis e não fundamentadas – que levam, depois, ao saudosismo, ao sonhar com o que poderia ter sido e não foi – e a ações desgastantes para o sistema. A avaliação não é, entretanto, conservadora; pressupõe um certo adualismo, exigência do próprio pensar educacional, mas o seu compromisso real é com a mudança e a transformação que somente a partir dela (avaliação) pode resultar. Avaliação nada tem a ver com a manutenção do *status quo*, a reprodução do pensamento acabado; a avaliação não gera um quadro de revolução, mas conduz a mudanças desejáveis, com as quais está comprometida.

## AVALIAÇÃO E REFLEXÃO

A avaliação está integrada no processo contínuo de refletir sobre as diversas áreas educacionais e seus problemas. A reflexão é constante e envolve o administrador, com os seus problemas para o bom funcionamento do pequeno/grande mundo que é a escola, o pessoal técnico – supervisores e coordenadores – que deve refletir sobre a diversidade dos problemas de seu dia a dia, envolvendo, inclusive, problemas de relacionamento humano, e o professor, supostamente um técnico em promover a aprendizagem, que precisa refletir como desenvolver habilidades e aptidões, além, naturalmente, de procurar fixar todo um conteúdo associado a um conjunto de experiências curriculares. A reflexão sobre essa diversidade mantém uma relação biunívoca com a avaliação; pois não se pode pensar em educação e ensino sem, simultaneamente, uma valoração de outros elementos diversos. A avaliação não é um conjunto de técnicas para o levantamento de informações sobre diferentes sujeitos, mas um momento permanente na reflexão sobre os problemas educacionais. A contradição momento/permanente justifica-se, porque se refere não a um instante temporal, mas a um elemento indispensável ao constante refletir sobre a educação e seus problemas.

Assim, todos, na comunidade educacional – administradores, técnicos e professores são avaliadores, na medida em que refletir sobre educação é avaliar o próprio problema.

## AVALIAÇÃO E POLÍTICAS PÚBLICAS

Se considerarmos o nosso contexto, observamos que nem sempre os programas de avaliação resultam de uma compreensão aprofundada das suas finalidades e do caráter impositivo que devem ter. A par dos modismos habituais, a avaliação é muitas vezes decidida de uma forma quase improvisada e passa a ser considerada prioridade. Um programa de avaliação deve resultar de uma definição de políticas que dimensionam um quadro específico com vistas à qualidade da educação. Não se avalia apenas para dizer que avaliou e que se possui um banco de dados computadorizado sobre múltiplos aspectos da educação. O refletir associado à avaliação decorre da necessidade de promover um ensino/educação que seja de qualidade. A própria reflexão mostra, por outro lado, que a avaliação deve ser de natureza diversificada, procurando fazer com que diferentes programas analisem e solucionem diferentes problemas. É preciso que compreendamos um programa de avaliação em seus fundamentos e objetivos, para que possamos definir e influenciar sobre as políticas públicas a serem estabelecidas e implementadas com seriedade, tendo em vista que a má qualidade da educação é catastrófica e a sociedade se vê solapada em um de seus elementos básicos, ficando na dependência de outras culturas mais bem estruturadas.

## AVALIAÇÃO E O AVALIADOR

Falamos, anteriormente, em reflexão, que resultaria numa análise que incluiria, também, um processo de avaliação e ressaltamos o papel de administradores, técnicos e professores, que, em princípio, estariam qualificados para esse desempenho. Aí incidimos em um ponto nuclear, pois a análise exigida é de natureza bastante complexa, não admitindo improvisações. É fundamental que o avaliador possua uma formação geral e que tenha vivenciado os problemas do ensino, especialmente através

da experiência de sala de aula. Vimos que a tarefa do avaliador exige habilidades específicas, desenvolvidas por intermédio de uma formação quantitativa, inicialmente, e qualitativa, posteriormente, conforme a orientação estabelecida por Robert Stake na Universidade de Illinois (Urbana). O desenvolvimento de capacidades quantitativas, apenas, pode levar a simplificações, às vezes quase esotéricas para os não-iniciados; a formação qualitativa, unicamente, por outro lado, leva a generalizações filosofantes, que pouco ou nada informam sobre a realidade para aqueles que devem tomar as decisões. Ambas as orientações completam-se e possibilitam ao avaliador explorar o problema em seus diferentes aspectos. Infelizmente, alguns exaltam apenas um tipo de formação, menosprezando o outro, o que demonstra um posicionamento equivocado, pois o quantitativo e o qualitativo se completam. Em algumas circunstâncias um impõe-se ao outro, sem que isso signifique que esse quadro seja o melhor, mas apenas o mais adequado à situação apresentada. A escolha da abordagem correta, associada à qualificação do avaliador, permite, desse modo, que falsas interpretações sejam evitadas e julgamentos igualmente falsos se concretizem.

## AVALIAÇÃO E CONTROVÉRSIAS

Alguns programas educacionais costumam dar margem a controvérsias. Pensamos, a título de ilustração, nos chamados programas de matemática moderna em oposição à matemática clássica.

Isso costuma gerar situações muitas vezes passionais e, portanto, emotivas, que dão margem ao surgimento de elementos que dificultam a tarefa do avaliador. É preciso, portanto, que o avaliador, na medida do possível, mantenha-se isento, sem tomar aprioristicamente um posicionamento. Isso ocorre com bastante frequência, ao contrário do que se possa pensar. É o resultado do desejo de provar que o novo é melhor, em oposição ao antigo, o que não é necessariamente verdade. A responsabilidade do avaliador, no caso, é bastante grande, exigindo que, graças à sua capacitação e à experiência adquirida ao longo da prática habitual, possa identificar essas situações desfigurantes e, assim, estabelecer meios adequados para esclarecer os problemas, sem envolvimentos ou controvérsias

que anulam o trabalho, fazendo com que não tenha validade e, *ipso facto*, não mereça credibilidade entre aqueles que usam na prática os dados que o avaliador lhes proporciona.

## AVALIAÇÃO E NOVAS METODOLOGIAS

O processo de avaliação ao longo do tempo precisa ser aprimorado, novas metodologias introduzidas e modificações realizadas, para que a avaliação não se transforme numa prática burocratizada, repetitiva e que acabe desmotivando a clientela a que ela se destina Isso é uma verdade inconteste e já começa a ser sentida em países que a usam intensamente, impondo-se um trabalho de motivação para que o descrédito não passe a dominar. Por outro lado, também, é preciso cautela, em relação a novidades não inteiramente testadas, e cujos resultados, além de nada informarem, não contribuem para o aprimoramento do processo de avaliação. Ainda que a situação seja ideal, isso não significa que não se concretize na realidade. É preciso que a avaliação apresente informações relevantes ao sistema e deixe bastante claro que os serviços que presta são realmente importantes, o que nem sempre é sentido. Poder-se-ia argumentar que estamos diante de uma situação ideal, nem sempre concretizável, considerando as forças negativas que existem dentro do sistema e que se opõem às avaliações, porquanto acabam por expô-las, denunciando a sua ação. Essa é uma das muitas finalidades da avaliação: identificar reações negativas e superá-las, apresentando caminhos de excelência para a melhora do sistema.

## AVALIAÇÃO E SEUS USUÁRIOS

Quem são – ou deveriam ser – os reais usuários das avaliações? As avaliações não se realizam para uso dos avaliadores, que são agentes na identificação dos problemas. Os administradores e técnicos devem ser os usuários dos resultados das avaliações e os implementadores do conjunto de ações ou de ações isoladas para a solução dos problemas. Entretanto, aqueles a quem as avaliações se destinam são realmente os professores, eles próprios avaliadores permanentes do próprio sistema e de tudo aquilo

que o aciona. Professores, entretanto, costumam alegar que os resultados das avaliações, especialmente as realizadas externamente por terceiros, não servem às suas necessidades e interesses. Por sua vez, os avaliadores queixam-se de que os resultados dos seus trabalhos não são utilizados devidamente, não só por professores, mas também por administradores e técnicos. Isso, sem dúvida, é uma realidade indiscutível. Áreas ministeriais e Secretarias de Estado possuem um número enorme de avaliações (e pesquisas, às vezes por solicitação oficial) e as mesmas repousam tranquilamente nos arquivos, ignoradas ou desconhecidas, não sendo utilizadas na definição de políticas e nem para a elaboração de projetos. Um descompasso realmente existe, e deve ser eliminado para que as avaliações resultem em ação interativa entre todos os membros que atuam no sistema educacional, o que somente se consegue com uma disseminação adequada dos elementos informativos às clientelas interessadas.

## AVALIAÇÃO E INFLUÊNCIA POLÍTICA

A interação entre os elementos ligados à avaliação mostra que o avaliador educacional, mesmo quando não pretende ter influência política sobre o sistema, acaba por exercê-la. E por que isso ocorre, independentemente da sua vontade? Pensemos, inicialmente, que a avaliação, assim como a pesquisa educacional, visa a coletar dados da realidade para autoconhecimento da sociedade (alunos, pais, professores, técnicos e administradores), inclusive daqueles que são responsáveis pela definição de políticas que mais tarde se vão transformar em projetos com vistas ao desdobramento de ações para a promoção da qualidade da educação. Ora, ainda que em termos teóricos, tudo deveria partir da avaliação e dos seus resultados, daí a responsabilidade do avaliador, que, dependendo da metodologia delineada e dos resultados obtidos, assim como de seu relatório, das suas conclusões e das propostas de ação pode encaminhar o sentido das decisões políticas a serem definidas e implementadas pelas autoridades educacionais. Nem sempre os avaliadores se dão conta da sua responsabilidade pública, das consequências de projetos inadequadamente propostos e da impropriedade de seus relatórios, acabando, dessa forma, por influir negativamente nas políticas

públicas, quando são consideradas. É importante que as avaliações sejam utilizadas pelo sistema, mas é sumamente necessário que essas avaliações sejam responsáveis, tendo em vista a possível influência que vão exercer.

## AVALIAÇÃO E CONSENSO

Uma única teoria da avaliação é o desejo não concretizável da maioria dos avaliadores. Assim como não existe apenas uma teoria da educação aceita pelo consenso – é impossível o consenso em educação, apesar das esperanças da técnica Delphi, que parte desse pressuposto –, não existe um consenso que também seria desejável em avaliação. As posições são muitas, às vezes antagônicas. Pode-se elaborar uma teoria com base em uma concepção de educação, mas será sempre uma entre muitas outras teorias igualmente possíveis. Algumas vezes, em educação, ficamos no domínio da idealidade. Elaborar uma teoria da avaliação que satisfaça a todos é difícil, praticamente impossível. As coisas tornam-se mais complexas quando se pensa em uma teoria que seja ao mesmo tempo uma teoria da interação política. lnteração política é um construto e admite diferentes posicionamentos difíceis de conciliar; por outro lado, uma teoria assim estruturada somente se completaria se a ela pudéssemos integrar uma teoria da determinação de fatores, pois, no fundo, o avaliador procura estabelecer os fatores determinantes que dimensionam o problema a solucionar. Acreditamos que o avaliador deva assumir uma posição filosofante, como pensador que é, e procurar basear-se em uma teoria (ou criar uma teoria) associada à interação política, não importando que vozes diversas se manifestem contrárias. O significativo é o seu trabalho e nesse quadro não podemos deixar de pensar em Tyler, que legou uma obra significativa, apesar de todas as resistências que enfrentou e soube superar.

## AVALIAÇÃO E EXPECTATIVAS

A avaliação afigura-se a muitos como o caminho da esperança, acreditando que, por intermédio da avaliação, se poderão obter

respostas inequívocas para os numerosos problemas que o campo educacional oferece. Sem dúvida, poderemos obter respostas para algumas questões, mas não para todas. Ao delinearmos um programa educacional, qualquer que seja, simples, sem maiores complexidades, sempre surgirão controvérsias. A avaliação do programa poderá fornecer certas informações sobre o seu possível mérito, indiscutivelmente, mas essas informações por mais completas que possam ser, por maior que seja a soma de dados coletados, nunca eliminará totalmente as controvérsias que sobre seus méritos possam existir. Poderá haver um certo convencimento, mas que não será suficiente para dirimir as dúvidas que possam existir. Neste caso, a avaliação que surge para alguns como a senda da esperança, pode vir a transformar-se em caminho do desapontamento. E por isso talvez seja árduo o *métier* de avaliador, que deve conviver com opostos: a esperança e a angústia da frustração. É preciso, pois, que o avaliador não parta de certezas absolutas, mas baseie o seu trabalho em possibilidades, em expectativas que possam vir a se concretizar, mas sem pensar que a tudo domina, que é o senhor de todos os saberes e que o seu domínio das tecnalidades seja capaz de eliminar as dúvidas que são parte da contingência humana.

## AVALIAÇÃO E PROCESSO POLÍTICO

A definição de um programa educacional insere-se em um conjunto maior de natureza social, que, por sua vez, resulta de um processo político, com vistas à transformação da educação e de sua agência responsável, a escola vemos, portanto, que há um encadeamento de momentos que se ligam e se interpenetram, constituindo um corpo uno. O avaliador muitas vezes é chamado a participar desse contexto e seu comportamento é de perplexidade. Seus modos, seus métodos de tratamento dos dados, suas sofisticadas estatísticas processadas por instrumentos de alta tecnologia permitem-lhe coletar e manipular grande soma de dados e, finalmente, chegar à esperada fase das conclusões. Conclusões que levarão a sugestões para eliminar situações ou criar um novo quadro. Nesse momento quase final do trabalho, o avaliador dá-se conta que o seu trabalho chegou a um impasse. Às conclusões opõe-se um elemento mais forte, o processo político

em que tudo está inserido. É um instante crítico, que demonstra a situação de impotência em que se encontra a avaliação e o avaliador educacional. Há razões outras que podem escapar à sua compreensão, mas que se impõem tendo em vista que a educação ocorre em um contexto político e parodiando o velho é sempre moderno filósofo francês, pode-se dizer que a política tem razões que a própria razão desconhece. Isso é verdade no mundo das emoções e no âmbito das realidades. A educação deve estar pronta para enfrentar essas realidades, sendo que isso não representa um demérito para o avaliador, quando as suas conclusões nem sempre são consideradas.

## AVALIAÇÃO E PESQUISA

A ideia de avaliação começa a dominar a nossa sociedade, graças, em parte, à atuação da mídia, chamando a atenção para alguns problemas que às vezes não são realmente prioritários, com o esquecimento de outros, mais relevantes e de repercussão social mais profunda. A questão da avaliação passou a dominar os mais diversos segmentos sociais, despertando o interesse de órgãos sindicais de um lado e no extremo oposto de grupos empresariais, ainda que por motivos diversos. Existe uma crença generalizada na necessidade de impor a avaliação como um procedimento normal ao sistema educacional. Percebe-se, entretanto, que essa ideia, aparentemente nova, nada mais é do que o desdobramento de uma ideia anterior, que tardiamente surgiu no âmbito do ensino superior: a prática da pesquisa. A pesquisa em educação é bastante recente, pelo menos em nosso contexto, e recebeu grande impulso na medida em que a pós graduação se consolidou nos principais centros de ensino superior. Uma nova atitude começou a formar-se em relação à análise dos problemas educacionais e um acúmulo de *know-how* tecnológico se estruturou, permitindo trabalhos empíricos. Avaliação e pesquisa seriam campos separados ou seriam na verdade áreas que se integram e se complementam? Tentamos, muitas vezes, delinear as características de uma e de outra, pontos de aproximação e de divergência; no entanto, afigura-se-nos que as atuais distinções são artificiais, são mais um exercício acadêmico, pois o evoluir de ambas as atividades

está fazendo com que a possível distinção que então existia esteja desaparecendo e a avaliação se transformando igualmente em uma forma de pesquisa. Podemos estabelecer como um imperativo atual que se formem pesquisadores para uma atuação eficiente na área da avaliação educacional.

## AVALIAÇÃO E *ACCOUNTABILITY*

A década de 60 assistiu a um *boom* da avaliação educacional no contexto norte-americano. Alguns milhares de projetos foram realizados com o investimento de centenas de milhões de dólares; ao final, alguns trabalhos se revelaram realmente relevantes e tiveram ampla repercussão, mas a maioria foi objeto de acerbas críticas, tendo em vista as limitações dos resultados apresentados. Aos poucos, começou a impor-se o conceito de *accountabillty*, palavra difícil de traduzir com exatidão; contudo, percebe-se que está relacionada à expressão *to be accountable for* – ser responsável por –, daí a tradução que usualmente empregamos responsabilidade educacional. Por outro lado, é preciso lembrar, também, que *account* tem um sentido contábil, financeiro. Haveria, assim, uma preocupação maior com as grandes despesas governamentais para obtenção de resultados muitas vezes decepcionantes, em decorrência, inclusive, da irrelevância metodológica da maioria dos projetos. Aprofundando-se o conceito de *accountability* sente-se que, além do aspecto financeiro, há uma grande preocupação com o que foi, com o passado, visando a definir méritos e a estabelecer a culpabilidade dos responsáveis pelos programas; parece-nos, entretanto, que a avaliação pode ser melhor utilizada na medida em que nos permita compreender os acontecimentos e a ter uma inteligência aprofundada dos processos educacionais, com o objetivo de fornecer subsídios para os novos educadores que se venham a envolver com futuros projetos e com outros programas, aproveitando os ensinamentos do passado e as experiências vivenciadas. A avaliação é um olhar para frente, um olhar em perspectiva, talvez a partir do que foi, mas sem querer culpabilizar pessoas ou instituições, bastando a angústia do possível insucesso. A avaliação guia; a avaliação não pune.

## AVALIAÇÃO E REALIDADES

O poeta diz com razão que navegar é preciso; diríamos que avaliar também é preciso, mas com as devidas cautelas, tendo uma bússola orientadora, a fim de evitar certos abrolhos e possíveis desapontamentos. Algumas vezes, arquitetam-se grandes planos, com vistas a grandes renovações e a esperança de promover uma revolução. Sonhar é igualmente necessário, mas dentro de certos limites, considerando as nossas realidades. Ideias são importadas sem que sofram um processo de aculturação; instituições são transplantadas com o esquecimento de que, atrás delas, há toda uma tradição acumulada ao longo dos tempos e experiências verificadas ao longo de um processo de validação. A situação não é abstrata, mas ocorrente em nosso contexto. O avaliador educacional vê-se muitas vezes envolvido nessa situação e é obrigado a opinar sobre propostas de renovação ou a elaborar planejamento que acompanhe a implementação dessas novas experiências. A sua situação pessoal é bastante crítica. Nem sempre pode recusar a sua participação; outras vezes, procurando agir com imparcialidade – na medida em que o ser humano pode ser imparcial – seus resultados contrariam as expectativas, sendo prudente, portanto, que o avaliador considere a possibilidade de insucesso do projeto de renovação e advirta os responsáveis pelo projeto. Sinceridade e honestidade são, obrigatoriamente, atributos de um avaliador consciente.

## AVALIAÇÃO E COMPROMETIMENTO

A atividade educacional implica intenso comprometimento dos envolvidos. Esta situação deve ser mantida ao longo de todo o desenrolar das ações programadas, inclusive com a participação ativa dos avaliadores, sejam estes especialistas externos ou professores, que, em última análise, são, realmente, os avaliadores dos programas nas escolas. A atividade educacional em países em transição socioeconômica é extremamente desgastante fisicamente e sem uma certa paixão, reflexo do comprometimento, fácil será cair na descrença, na prática do fazer pelo fazer, simplesmente. O avaliador, e não somente ele, mas todos os educadores, devem superar seus momentos de descrença e dúvida – fé e dúvida coexistem e a dúvida robustece a fé –, que podem levar ao negativismo e à in-

diferença, estados emocionais bastante frequentes em educação. A avaliação tem seus paradoxos – sem paixão, sem comprometimento é impossível uma análise racional. Isto é aparentemente contraditório – paixão e racionalidade, mas, na verdade, se não houver da parte do avaliador uma certa paixão que o leve a persistir, a manter-se acima e além das adversidades, que são bastante comuns, não terá a suficiente clareza de espírito para empreender uma análise racional que o trabalho quase artesanal do avaliador exige. Desse modo, o comprometimento, que é essencialmente político, leva o avaliador a agir (e a reagir) com racionalidade.

## AVALIAÇÃO E CONTROLE

Há alguns parágrafos consideramos a problemática da *accountability*. Voltemos a ela, porque esse conceito que surgiu há mais de trinta anos no contexto norte-americano começou a expatriar-se, passando a fazer parte do jargão tecnológico das avaliações nacionais. Será que essa ideia se aplica à nossa situação, que somente agora, nos últimos anos, menos de um decênio, começa a se preocupar com a avaliação, seus problemas e as inúmeras implicações que dela decorrem? Acreditamos que sim, considerando que havendo recursos, pessoal qualificado, planejamento adequado e ação implementadora consistente, os programas educacionais, quaisquer que sejam, serão bem sucedidos. Essa é a ideia mais explícita do conceito de *accountability*. Outras percepções também existem, como o seu conteúdo financeiro, entretanto, preocupa-nos um outro aspecto, que nos parece extremamente grave. Se necessitamos desse conceito a ponto de importá-lo e incorporá-lo à própria educação, não dizemos que apenas a avaliação o incorporou. A educação em geral sentiu necessidade de utilizar essa ideia, não unicamente a avaliação, e isso é preocupante, pelo menos para nós, pois revela que o sistema se apercebe que um quadro patológico se desenvolve no sistema, e necessita de elementos de controle, a fim de evitar que ele se desenvolva e comprometa a sua própria razão de ser, inclusive a avaliação, que passa a sofrer os mesmos malefícios, via todo um processo de metástase social. A avaliação, ao coletar e analisar dados, passa, assim, a exercer as funções de um elemento de controle da salubridade do sistema educacional.

## AVALIAÇÃO E COISA PÚBLICA

A informação é uma fonte de poder e deste modo faz com que muitos guardem elementos de informação, às vezes como justificativa para a própria existência. Isso refletiria a ideia de uma comunidade fechada em que somente as autoridades teriam acesso ao conhecimento dos dados. A avaliação opor-se-á a esta situação. Os dados da avaliação, obtidos através de uma *expertise* adequada não são propriedade exclusiva do avaliador, não pertencem à escola, não são de uso restrito dos autores e implementadores de programas, não pertencem às Secretarias de Educação e nem mesmo à autoridade centralizadora do Ministério da Educação, agente do Estado. Os resultados de uma avaliação constituem *res publica*, coisa pública, para uso da comunidade e conhecimento da sociedade. Assim, a avaliação não tem sentido em uma sociedade fechada, centralizadora e autoritária. A sua razão de ser está em uma sociedade aberta, descentralizada e democrática. A avaliação somente se pode considerar plenamente realizada na medida em que a utilização de seus resultados demonstre que não há um círculo fechado, que estamos em uma sociedade transparente que procura se conhecer e ter a suficiente coragem de apresentar suas deficiências e indicar os aspectos em que é bem-sucedida.

## AVALIAÇÃO E CREDIBILIDADE

A eficiência de um governo não está no fazer, mas no saber fazer em correspondência com as realidades vividas no contexto da sociedade. Ao querer fazer, sem antes consultar a sociedade e ouvir seus posicionamentos, um governo democrático, constituído por pessoas de formação democrática, mas ás vezes sem experiência democrática, acaba por revelar uma face autoritária até então desconhecida. A centralização na área educacional leva muitas vezes a situações contraditórias. Ao impor, por exemplo, a avaliação – processo fundamental no âmbito da educação, não temos dúvida, por uma questão de coerência com o próprio viver – sentimos um princípio de violência, que não é desejável. A avaliação deve resultar de uma aspiração própria e do desejo de autoconhecer-se, ao mesmo tempo que identifica seus problemas, seus êxitos e seus fracassos, ainda que essa colocação possa parecer utópica, mas sem uma utopia não se vive, nem sobrevive.

Avaliação em que o governo avalia seus próprios procedimentos não merece credibilidade. Falta-lhe credibilidade por querer impor uma avaliação não inteiramente independente, porque realizada por órgãos de certa forma ligados ao poder central ou que dele dependem para sobreviver. Cairíamos, dessa forma, no mundo inconsequente de verdades paradoxais (uma contradição, evidentemente), ou seja, em educação o suposto racionalismo de um governo centralizado pode levar ao autoritarismo, ao generalizar medidas possivelmente aceitas pela sociedade, desde que tivesse sido ouvida e os temas discutidos por suas instituições mais representativas.

## AVALIAÇÃO E OBSTÁCULOS

O avaliador, assim como o homem de Ortega y Gassett, é ele próprio e suas circunstâncias. Possui suas limitações e sofre restrições que lhe são impostas pelo contexto em que atua. Tem o domínio das técnicas, mas nem sempre as pode usar, tendo em vista os condicionamentos do público a que se destinam, ou, então, as empregam usando meios palatáveis de comunicação, a fim de que seu trabalho possa ser compreendido e utilizado. Nem sempre pode realizar o que, teoricamente, gostaria de concretizar, porque possui limitações orçamentárias, às vezes bastante sérias. Sendo o seu trabalho o resultado do esforço de uma equipe, em muitos casos não dispõe de pessoal qualificado ou deve agir com apoio de um grupo bastante restrito. Ao desenvolver o projeto não terá condições de superar muitas dificuldades e ele próprio em seu relatório deverá honestamente explicitar as dificuldades encontradas, as limitações enfrentadas e os aspectos que não puderam ser abordados, por razões diversas, e as soluções que não forem possíveis. O avaliador, assim como o pesquisador, e ele próprio é também um pesquisador, não é um super-homem que vence todos os obstáculos, enfrenta todas as situações e soluciona todos os problemas, armado com os instrumentos que lhe oferece a tecnologia de ponta. O avaliador é, em grande parte, o gerenciador de um programa, que, com diligência, procura torná-lo o mais eficiente possível, sem o uso de superpoderes que não possui. Seria uma ilusão pensar ao contrário.

## AVALIAÇÃO E GERENCIAMENTO

Chegamos a um ponto crítico nestas reflexões sobre a avaliação educacional. Perguntamo-nos: por que o sucesso do *Eight-Year Study*, de Tyler, ou o de Stufflebeam, baseado no modelo Contexto, lnput, Processo e Produto (CIPP)? É evidente que, primeiramente, havia a vontade política de realização das avaliações, sem o que nada é possível realizar. Havia, ainda, um modelo com sólida fundamentação teórica, independentemente das restrições que alguns elementos da comunidade de avaliadores apresentaram a essas formulações. Os projetos possuíam dotações orçamentárias necessárias à sua concretização, com verbas de fontes diversas. Isso tudo, entretanto, não seria suficiente para explicar o êxito desses projetos ou de quaisquer outros programas de avaliação; há um outro elemento de capital importância: o responsável pelo gerenciamento dos projetos. Tyler e Stufflebeam tinham experiência de ensino; participaram de atividades acadêmicas, inclusive administrativas, em grandes universidades, e possuíam a vivência de outros programas anteriormente executados. Tinham capacidade de liderança, aglutinando em torno de si instituições diversas e diferentes pessoas com formação variada; faziam valer, democraticamente, o seu comando, solucionando os vários conflitos de relações humanas; além de se imporem pelo domínio da área substantiva em que estavam envolvidos, ou seja, o gerente não é um burocrata, mas um cientista com experiências diversificadas. Em síntese, o êxito de qualquer programa de avaliação depende de um controle firme da parte de seu gerente e da sua capacidade de liderança.

## AVALIAÇÃO E LIDERANÇA DEMOCRÁTICA

Um programa de avaliação envolvendo todo um sistema educacional, ainda que realizado por amostragem, demanda negociações e acomodações, que são próprias de uma liderança democrática. A imposição de um programa ao sistema exige um preparo anterior, para atenuar as reações contrárias, que são inevitáveis. Qualquer avaliação mexe com o sistema e reações adversas devem ser esperadas. São parte da mecânica do processo, sendo necessário um trabalho inicial de acomodação para a diluição das forças adversas. Sem colaboração não é possível realizar esse

trabalho, que deve ser desmembrado em diferentes níveis, com várias linhas de coordenação e supervisão. O poder numa avaliação, ainda que haja um gerenciamento geral, deve ser diluído e as responsabilidades compartilhadas em diferentes momentos e instâncias. Seria uma ideia platônica admitir-se que, definido um programa, o mesmo passará a ser aceito sem questionamentos, com a concordância integral dos vários integrantes do processo a ser desenvolvido. É necessário um trabalho prévio, em que fique demonstrada a decisão política de realizar o programa, e sejam esclarecidos os problemas suscitados pelo próprio sistema, numa tentativa, aliás compreensível, de autodefesa. Defesa em relação a algo que desconhece e que precisa ser definido para que seja compreendido o seu significado, a sua importância, o uso de seus resultados e as modificações que possivelmente serão realizadas, como decorrência da avaliação. É importante que se saiba por que está sendo feita a avaliação, suas finalidades e que providências serão tomadas no futuro. Avaliar por avaliar, simplesmente, não faz sentido, é necessário que haja consequências decorrentes da avaliação.

## AVALIAÇÃO E AGÊNCIAS FINANCEIRAS

O avaliador, no desempenho do seu trabalho, vê-se envolvido em diferentes contextos, compreendendo administração (e burocracia), quadros técnicos, corpo discente e docente, além da estrutura financeira que lhe fornecerá os meios para desenvolver o projeto. É bastante frequente que essa estrutura financeira seja representada por instituições nacionais e internacionais, que têm a sua cultura própria, o seu corpo técnico especializado, mas não necessariamente em avaliação, e que por intermédio dos projetos, fazendo exigências, tentem interferir nas políticas públicas. Acreditamos que a parceria com agências financeiras seja desejável e até mesmo bem-vinda; contudo, o avaliador precisa acautelar-se em relação a diferentes aspectos, inclusive adotando uma atitude agressiva, se necessário for, e abandonando posições de prudência política. É preciso cautela em face da tentativa de impor modelos que correspondem a outros contextos e que nada têm a ver com a cultura educacional. O modelo a ser usado deve resultar de um consenso entre as partes atuantes no

processo de avaliação, com envolvimento de até mesmo as agências, mas sem a prevalência de suas ideias, muitas vezes exóticas. A imposição de consultores é preciso ser vista igualmente com reserva, pois quase sempre apresentam extenso currículo, mas desconhecem a realidade nacional e agem em termos inteiramente abstratos, sugerindo metodologias que não atendem às necessidades do projeto e nem aos seus objetivos. Ainda, frequentemente, propõem técnicas de tratamento para fins de futuras pesquisas que não se ajustam à natureza dos dados coletados e exigem dos participantes da avaliação, especialmente em nível de corpo docente, uma familiaridade com elementos que não fazem parte da sua cultura pedagógica. O avaliador deve fazer frente a essas situações, impondo-se, exercendo a sua função de avaliador, dentro das suas responsabilidades. É evidente que nem sempre se pode prescindir da colaboração das agências financiadoras. É preciso que o responsável pela avaliação não seja um "guardião platônico" do seu projeto, mas adote uma posição que resguarde o seu planejamento, sem, entretanto, criar uma situação conflitiva que impeça a concretização do trabalho.

## AVALIAÇÃO E META-AVALIAÇÃO

A experiência pessoal nos tem demonstrado que, qualquer que seja o processo de avaliação, o sistema tende a permanecer imóvel, sem promover iniciativas que levem à atuação do avaliador. É preciso que sua iniciativa pessoal atue, coordenando as várias atividades e promovendo o envolvimento de todos os participantes. Apresentados os resultados, muito possivelmente pouco se fará em termos de modificações, salvo se estas vierem através de sugestões no próprio relatório do avaliador. É lógico que existem exceções; no entanto, considerando os interesses políticos em entrechoque na área educacional, talvez seja necessária uma intervenção do avaliador. As coisas facilmente se acomodam e tendem a permanecer como se encontram, prevalecendo o estado de inércia. Assim, obtidos os resultados do trabalho é necessário que se promova a avaliação dos mesmos, que, em outras palavras, se faça uma análise de todo o processo, ou seja, que haja uma meta-avaliação (avaliação da avaliação) a fim de que seja julgado o seu valor. A avaliação da avaliação, segundo determinados padrões

(Stufflebeam), deve ser externa, na medida do possível, considerando o envolvimento do avaliador e o seu comprometimento, o que pode gerar vieses nos julgamentos, ficando comprometidos os juízos que se fizerem. Além das características já apresentadas, o avaliador também deve ter iniciativa e senso de oportunidade, para agir no momento próprio.

## AVALIAÇÃO E INTERESSES CONFLITANTES

A avaliação exige um processo de negociação com diferentes segmentos que muitas vezes apresentam interesses conflitantes. Vejamos, por exemplo, os casos da avaliação institucional ou da avaliação de cursos de 3º Grau, após a conclusão do bacharelado, ideias que começam a se concretizar em nosso contexto educacional. É possível que a maioria da sociedade esteja em princípio de acordo com as ideias básicas das propostas, como no nosso caso pessoal; no entanto, especialmente a comunidade acadêmica apresentou reações negativas, algumas de origem corporativa, tendo em vista que não participou do seu planejamento e não discutiu nenhuma proposta, sendo surpreendida por medidas impositivas. Assim, a avaliação deve disseminar informações que diluam reações contrárias, que atenuem as arestas apresentadas, que muitas vezes invalidam o trabalho de avaliação. É mais do que evidente que a avaliação deve apresentar informações que possibilitem a tomada de todo um conjunto de decisões corretas; entretanto, essa mesma avaliação deve fazer um levantamento exaustivo dos elementos de informação para que sejam apresentados à sociedade no processo de negociação com vistas à sua viabilidade.

## AVALIAÇÃO E SUA DINÂMICA

O trabalho de avaliação faz-se de forma progressiva e à medida que avança exige constantes adaptações. O planejamento, por mais bem feito que seja, considera o máximo de variáveis possíveis, mas não tem condições de prever o imponderável; desse modo, a avaliação precisa ser feita com suficiente plasticidade para que possa enfrentar esses fatores imprevisíveis de forma efetiva e bem-

-sucedida. A avaliação de um sistema educacional em nível de Estado, por exemplo, apresenta inúmeros problemas que não são antecipáveis, acontecem na medida em que o trabalho se processa; por outro lado, ainda que os treinamentos realizados em diferentes níveis sejam eficientes, as informações passadas dentro do sistema podem gerar diferentes procedimentos, às vezes com inovações, e comportamentos não previstos que podem afetar todo o processo, impondo-se, desse modo, adaptações para que o planejamento como um todo não seja afetado e os objetivos estabelecidos desvirtuados. A avaliação, assim, exige toda uma dinâmica especial, reflexo de sua plasticidade e capacidade de interação nas várias situações que podem afetar o trabalho do avaliador.

## AVALIAÇÃO E TOMADA DE DECISÃO

A avaliação, qualquer que seja o seu objetivo, pretende chegar a um ponto que permita a tomada de decisão, às vezes pelo próprio avaliador, mas na maioria das vezes pela pessoa ou instituição interessada no trabalho. A decisão deve emergir da própria avaliação, não é algo externo, uma simples consequência do projeto, mas uma decorrência do que foi trabalhado. O que desejamos destacar é que a decisão não é um fruto exclusivo do avaliador, mas uma emergência de todo o processo. Tudo na avaliação deve levar a essa emergência, que justifica o trabalho, e dá sentido ao que foi realizado.

## AVALIAÇÃO E DÚVIDAS

Ao longo de um trabalho de avaliação as dúvidas, as inseguranças e as incertezas são constantes e devem ser superadas. O avaliador deve ter uma estrutura mental capaz de superar esse quadro depressivo que se forma e que o leva a perceber que muitas vezes está sozinho e que por si deve vencer suas próprias angústias. A situação do avaliador se enquadra nesse contexto, apesar de possuir uma equipe que o auxilia e estar envolvido por numerosas pessoas que o observam e esperam seus resultados, mas na hora da tomada de decisão sobre como agir em relação a determinado aspecto, ele é um solitário na multidão, e deve agir por si. Então,

o agir do avaliador deve resultar da plausibilidade de suas ações que não devem conflitar com os interesses e as expectativas políticas da sociedade em que atua. Avaliar não é estabelecer um confronto, mas agir sensatamente para que os resultados dessa avaliação sejam plenamente aceitos pela sociedade.

## AVALIAÇÃO E SEUS RESULTADOS

A avaliação tem objetivos definidos, mas o fato de não os alcançar não significa que tenha fracassado. Isso é da natureza do próprio processo. Às vezes, apenas alguns poucos objetivos são alcançados, o que já representa uma vitória. O mesmo se pode dizer em relação aos resultados, que nem sempre são positivos, refletindo o sucesso de um programa ou o êxito de uma determinada inovação educacional. O insucesso fornece igualmente informações, que são importantes nas decisões. Nem todos os sonhos se concretizam, mas o esforço para realizá-los é válido, e por si próprio o empenho de concretizá-los já constitui uma recompensa. Assim, também, na avaliação. Apesar de nem sempre chegarmos a resultados favoráveis, isso não significa fracasso, antes, é uma vitória, mostrou que o programa avaliado não funcionou e deve ser modificado ou substituído, que o material arduamente construído não satisfez plenamente e deve ser recuperado em certos pontos ou descartado na sua totalidade.

## AVALIAÇÃO E DEFINIÇÃO DE OBJETIVOS

A questão dos objetivos em avaliação é às vezes colocada em termos retóricos; ou melhor, enfatizam-se os objetivos específicos, que são de natureza operacional e sem os quais nem sempre se pode construir o instrumental necessário ao trabalho. Ainda que possa existir uma avaliação *goal-free*, é indiscutível que, independentemente de posicionamentos ideológicos, a definição de objetivos é importante, por serem orientadores das várias atividades. Isso não significa que a avaliação deve ficar restrita a objetivos operacionais e/ou comportamentais. A avaliação possui objetivos amplos, muitas vezes de grande abstração, e nem sempre mensuráveis, quantificáveis e até mesmo nem sempre observáveis de

imediato. A avaliação visa a provocar um impacto, procura despertar interesses, tenta gerar atitudes positivas; a avaliação, em síntese, colabora para a promoção da qualidade da educação. O que é qualidade em educação? É um atributo que sabemos existir – um construto –, sobre o qual hipotetizamos, porque acreditamos no seu existir, mas dificilmente conseguiremos defini-lo, e nem mesmo há necessidade dessa definição, para sermos concretos.

## AVALIAÇÃO E VIESES DO AVALIADOR

A educação é uma arte, a educação é uma ciência. Diríamos, ainda, que a educação é uma área de incertezas, resultantes do entrechoque de opiniões conflitivas, não se conseguindo muitas vezes o consenso esperado. O pensamento dos educadores é sofisticado e nem sempre acessível aos não-iniciados. Talvez haja uma certa intencionalidade nesse comportamento, cujos objetivos são às vezes nebulosos. No entanto, é exigido do avaliador que ele seja claro, exato, preciso, objetivo, positivo e explícito nas suas palavras e ações, o que nem sempre é possível, porque todo avaliador possui vieses, que resultam da sua formação, das suas predileções, das suas experiências profissionais e das suas idiossincrasias ideológicas. O avaliador não é um ser perfeito, possui suas deficiências e prejuízos, próprios de sua condição humana, que deve ser respeitada. E muitas vezes o avaliador deixa de ser preciso e cai na subjetividade; contudo, há um espaço para a subjetividade na avaliação, mesmo quando trata com elementos objetivos. O homem – avaliador – não consegue fugir à sua condição humana e, assim, a subjetividade faz sentido no seu existir profissional.

## AVALIAÇÃO E QUESTÕES SIMPLISTAS

A responsabilidade do avaliador é enorme na condução do seu trabalho, pois a comunidade – escola, alunos, pais, professores – se volta para ele, que, assim, se torna o ponto focal que, supostamente, deve iluminar todos os problemas de interesse dos diversos segmentos. O que ele faz pode vir a ser o norte, o objetivo maior de todo o sistema, que passa, então, a atuar no sentido de prever uma imagem de suposta eficiência do ensino, dos pro-

gramas, dos materiais e de inovações que por ventura estejam sendo avaliados. Assim, deve ponderar as consequências do seu agir, que estará influenciando o comportamento de toda uma constelação de interessados no seu trabalho.

Apesar desse interesse pelos resultados da avaliação, sobretudo se há possibilidades de que sejam favoráveis ao objeto avaliado, o sistema não apresenta questões desafiadoras ao avaliador, permanece entregue ao seu quietismo burocrático, à espera de que as coisas aconteçam por obra do trabalho de avaliação. O sistema educacional, salvo as exceções de praxe, como convém acentuar, atua por inércia, não se autoprovoca, não se estimula, daí, como anteriormente foi registrado, o trabalho de avaliação ser considerado um ponto nuclear, a partir do qual novas perspectivas se abrirão, o que não deixa de ser parcialmente correto. A expectativa, entretanto, é que ao sistema caiba a iniciativa de ações, a partir de perguntas estimulantes para o avaliador, que quase nunca são apresentadas, ficando o avaliador como responsável pela sua propositura e, igualmente, pelas respostas desejadas pelos usuários da avaliação.

O problema da avaliação é muitas vezes proposto ao avaliador de forma bastante simplista: avaliar o sistema; avaliar o livro; avaliar o material didático etc. A questão é inteiramente vaga, nada diz ao avaliador, não é explícita, cabendo, então, ao avaliador determinar o que se deseja, usar sua intuição, e com sensibilidade determinar os objetivos da avaliação, os problemas envolvidos e a destinação que será dada aos elementos de informação. A avaliação não é um agir individualista, mas um trabalho solidário, que demanda o envolvimento da comunidade em que se situa o objeto da avaliação. As afirmações devem ser diretivas a fim de que perguntas adequadas sejam propostas para, ao fim e ao cabo, termos dados de informação que proporcionem uma possível solução para a indagação inicial.

## AVALIAÇÃO E SEU PROJETO

Avaliar não é simplesmente construir instrumentos e levantar o máximo possível de dados informativos. Às vezes, o excesso de dados acaba por prejudicar o trabalho, porque resultam de questionamentos inapropriados. Antes, portanto, de planejar ade-

quadamente a coleta de dados, é fundamental, diríamos mesmo que indispensável, a total imersão do avaliador no projeto, para conhecer a sua filosofia, os seus propósitos, as suas expectativas e toda a sua linha de concepção. A identificação com todos os detalhes do projeto representa mais do que o passo inicial para o sucesso da tarefa proposta; somente depois dessa interação projeto/avaliador é que o projeto passa a realmente existir, adquire vida e é capaz de produzir frutos. Talvez as primeiras perguntas a serem feitas sejam: como esse projeto foi efetivamente concebido? quais as suas motivações? quais os seus reais propósitos? Aí, sim, podemos começar a interagir – programa e avaliador – em busca de respostas aos problemas expostos.

As perguntas que fazemos em avaliação muitas vezes têm múltiplas respostas, considerando-se que há diferentes interessados – sociedade e escola –, que esperam respostas diferenciadas. A avaliação não dá uma única resposta, às vezes nem mesmo oferece uma resposta, mas apresenta outras indagações, que são igualmente importantes para a construção de conhecimento aprofundado do objeto. A indagação gera outras interrogações, que conduzem a novas situações ou até mesmo a outros problemas, e com isso o conhecimento se acumula, novas perspectivas se abrem e a avaliação passa a ter validade.

É preciso cautela no levantamento de dados informativos, que podem gerar um quadro enganador. A preocupação muitas vezes está no acumular o máximo possível de dados, mas isso não é suficiente. Ainda que haja toda a probidade possível da parte do avaliador, se as perguntas não forem aquelas que realmente devem ser feitas e a qualidade das respostas garantida, o simples acúmulo de elementos de informação não faz sentido. É preciso que toda avaliação tenha pelo menos duas audiências: a dos que aplaudem e a dos que criticam; dessa forma, as perguntas devem apresentar fatos úteis, que justifiquem os aplausos e possam responder àqueles que por ventura venham a apresentar suas críticas, muitas vezes bastante procedentes.

## AVALIAÇÃO E OCORRÊNCIAS PROVÁVEIS

Quando começa efetivamente uma avaliação? É uma pergunta que nos temos feito reiteradas vezes e a resposta foge ao comu-

mente apresentado: o seu começo está com o planejamento, com o *design*. Parece-nos que, antes disso, antes de planejar, o avaliador deve ter uma intimidade maior com o objeto a ser avaliado, para, depois, então, iniciar o seu projeto. Talvez começar pela análise de experiências similares já realizadas, pois nada de novo existe sob o sol; o que pretendemos realizar às vezes já foi feito com eficiência, e é preciso aproveitar experiências anteriores. As avaliações têm uma base de sustentação teórica, sendo útil fazer uma *review* do embasamento de projetos semelhantes anteriormente realizados. É preciso considerar que a avaliação será objeto de exame crítico pela comunidade e, assim, antecipadamente, deve ponderar a respeito dos pontos vulneráveis a críticas. Há todo um conjunto de expectativas com relação ao projeto que precisa também ser analisado, pois muito possivelmente várias dessas expectativas não sejam concretizáveis, por circunstâncias do momento, e é necessário evitar frustrações. Antes de planejar é necessário pensar nos possíveis usuários dos dados e na destinação que será dada aos resultados. Após estas considerações, faz sentido iniciar o planejamento.

A precisão dos resultados em uma avaliação é possível e necessária. Se não houver precisão, não terá credibilidade, daí sua imperiosa necessidade. A possibilidade de obter resultados precisos exige um trabalho piloto, ou seja, um trabalho que permita validar a forma como a avaliação será conduzida. Seria quase uma pré-testagem de todo o processo para eliminar os elementos comprometedores da confiabilidade dos resultados. Vendo as coisas em função do nosso contexto, sente-se que isso nem sempre é considerado. Há um processo de improvisação acelerada, na falsa crença de que uma avaliação possa ser realizada *ex-abrupto*. É forçoso reconhecer que precisamos encarar com cautela muitos resultados ufanisticamente apresentados, pois podem estar comprometidos na sua infraestrutura.

A possibilidade da realização de um protótipo precisa ser considerada. Iniciar uma avaliação sem ter perfeita consciência do que vai acontecer, dos problemas que poderão surgir, comprometendo todo o processo, é temerário e um avaliador com experiência terá certamente o bom senso de evitar essa situação. A pré-realização de uma avaliação permitirá que o avaliador considere uma amplitude bastante grande de ocorrên-

cias prováveis que na prática do dia a dia podem efetivamente surgir, às vezes de forma inesperada e sem possibilidade de uma solução imediata.

## AVALIAÇÃO E REPRODUÇÃO DE MODELOS

A questão dos planos de avaliação precisa ser considerada com bastante atenção, tendo em vista que em muitos casos são pouco flexíveis, preocupados com minúcias, apresentando uma formatação estereotipada e um conteúdo bastante repetitivo. Falta a esses planos uma personalidade própria, que lhes dê identidade, uma característica pessoal. Há casos em que os planos se repetem *ipsis verbis* de instituição para instituição, de programa para programa, sem considerar a diversidade dos contextos educacionais e das características culturais. Há um certo mecanicismo, uma rigidez nessas planificações. É preciso que os projetos de avaliação fujam a esquemas pré-definidos e que considerem a diversidade das situações, o que nem sempre ocorre.

O avaliador deve valorizar o seu trabalho, recusando projetos que não ofereçam oportunidades de uma atividade criativa. Explicitando a questão, podemos dizer que muitos projetos são parte da burocracia educacional, que é ponderável e adota esquemas que não dão margem a novas indagações, a novos procedimentos. Em síntese, nada acrescentam ao avaliador e nem concorrem para o aprimoramento do próprio processo de avaliação. Repetem-se *ad nauseam*, são simples reproduções de outros modelos, nem sempre escolhidos adequadamente. É um trabalho que leva à exaustão intelectual e os resultados não são compensadores, pois não fertilizam a área educacional, não geram novos conhecimentos. Repetem procedimentos sem nada acrescentar. A impossibilidade de criatividade é motivo suficiente para a recusa de um projeto de avaliação, assim como a falta de liberdade. Se o avaliador não tiver liberdade de atuação, se não puder agir com autonomia (relativa em alguns casos, mas absoluta em outros), segundo sua capacitação e aceitando a responsabilidade por suas ações, é melhor que não inicie o projeto, porquanto se sentirá cerceado no seu agir. O avaliador precisa convencer-se de que suas atividades sofrem algumas limitações e com estas deve conformar-se. Ainda que deseje divulgar os dados

de forma que possam ser amplamente utilizados, é preciso reconhecer que a avaliação não é sua propriedade particular, devendo, assim, conciliar os seus desejos com os interesses da outra parte (agências financiadoras, órgãos governamentais, instituições etc.). Por outro lado, precisa considerar que crenças e convicções não devem influenciar a avaliação, produzindo um trabalho ideologicamente contaminado, o que muitas vezes ocorre; além disso, o avaliador não deve insistir em permanecer indefinidamente em um mesmo projeto, apropriando-se dele como algo exclusivamente seu. O rodízio do avaliador por vários projetos é salutar para a avaliação e especialmente para o próprio avaliador.

## AVALIAÇÃO E ESTRATÉGIAS DE DIVULGAÇÃO

As relações entre avaliadores e administradores muitas vezes são tensas. O avaliador nem sempre considera aspectos que o executivo acredita fundamentais. Por exemplo, custos. E criam-se alguns conflitos. O administrador, por sua vez, quer o domínio das informações, mantê-las sob seu controle, pois parte do pressuposto de que ter informações é ter poder. O avaliador julga que esses dados devam ser amplamente divulgados para diferentes segmentos a fim de que os frutos do seu trabalho sejam disseminados e produzam efeitos.

Nada mais legítimo, especialmente na área da avaliação educacional. Entretanto, é necessário um entendimento entre as partes. A avaliação depende em muitos aspectos de apoio logístico, que a administração proporciona; por sua vez, o executivo necessita dos dados da avaliação para a tomada de decisões. Ambos, portanto, completam-se e devem respeitar suas respectivas áreas de atuação.

Há uma preocupação legítima dos avaliadores com os aspectos éticos da divulgação dos resultados. A forma como essa divulgação é feita pode significar o descrédito do projeto e, consequentemente, a sua repercussão negativa e o menosprezo dos resultados. Há que adotar, portanto, medidas acautelatórias para divulgação dos dados ao final do estudo, para que os mesmos não sejam invalidados. Outro aspecto a considerar refere-se à privacidade dos dados, que não pode ser descurada. Ainda que em muitos lugares, como, por exemplo, na Inglaterra, os resul-

tados de avaliações educacionais sejam apresentados com a individualização das escolas participantes, acreditamos seja questionável esse comportamento, tendo em vista as suas diferentes implicações. Além disso, em muitos casos, existem obrigações contratuais, que limitam a divulgação parcial ou total dos dados, sendo um aspecto a considerar pelo avaliador.

A estratégia de divulgação dos resultados de avaliações precisa ser considerada com cautela. Qual o procedimento a seguir? Resultados parcelados ou apenas ao final dos trabalhos? A divulgação dos resultados globais, procedimento mais frequente, costuma causar um grande impacto, mas, simultaneamente, provoca interpretações que podem ser distorcidas, especialmente pelos órgãos da mídia, que nem sempre estão interessados no fato científico, mas na repercussão que terá junto ao público. A divulgação parcelada parece-nos a mais sensata, porque o público interessado passa a participar das várias fases do trabalho e assim o acompanha até o final, discutindo seus aspectos e vivenciando os problemas em diferentes momentos.

## AVALIAÇÃO E RELAÇÕES HUMANAS

O fator pessoal na avaliação é um elemento que não pode deixar de ser considerado, inclusive tendo em vista a possibilidade do surgimento de situações conflitivas na própria equipe, já que a avaliação é uma obra realizada em conjunto. Há, pois, na avaliação, um componente de relações humanas que precisa ser gerenciado com o objetivo de evitar quadros de conflito entre os membros do grupo de trabalho. Uma equipe em crise de relacionamento não consegue produzir, desgasta-se e fragmenta-se, inviabilizando o trabalho. Outro problema humano centra-se no interesse das autoridades em relação aos dados da avaliação. Este interesse algumas vezes somente surge após a divulgação dos resultados e do impacto que por ventura os mesmos provocaram. É preciso que as autoridades, inclusive as administrativas, na área da educação, aprendam com os dados e atentem para a sua relevância e significado. O fator pessoal também está presente no avaliador, no seu desejo de mostrar o que conseguiu positivar e as conclusões a que chegou, havendo necessidade de um controle do que poderá dar

margem a exibicionismos e a procedimentos inconsequentes. Ao avaliador deve-se exigir uma postura serena de cientista.

O equilíbrio na comunicação dos resultados de uma avaliação é indispensável, sem o que o trabalho poderá resultar prejudicado. A carência de elementos informativos é um obstáculo para uma boa avaliação, mas, por outro lado, um número excessivo de informações, muitas vezes conflitivas, não permite, igualmente, uma análise razoável das variáveis envolvidas. É necessário, portanto, que no delineamento da avaliação seja feita uma ampla discussão sobre os dados a coletar, não incidindo no erro frequente de levantar dados simplesmente por levantar um maior número de elementos sem saber, muitas vezes, qual a destinação que será dada a esses elementos. A avaliação não pode conter no seu âmago elementos que a invalidem.

O problema da comunicação dos resultados de uma avaliação reveste-se da maior complexidade, e exige reflexão do avaliador. As comunicações formais, ainda que necessárias, devem ser suplementadas por outros meios informais, que às vezes esclarecem melhor, e de forma definitiva, as dúvidas surgidas, as incompreensões que muitas vezes decorrem de uma leitura imperfeita dos resultados apresentados formalmente, por intermédio de relatórios técnicos, que nem sempre estão ao alcance da compreensão de pais, administradores e mesmo de professores, por excesso de tecnalidades. Acreditamos, por isso, que a comunicação deva ser feita em dois níveis; um, formal, técnico, para os especialistas; outro, simples, informal com caráter de divulgação, para acesso de um público mais amplo.

## AVALIAÇÃO E REPERCUSSÕES

Ao fazermos uma avaliação precisamos pensar que os nossos procedimentos terão implicações nas próximas avaliações. Ou seja, ao avaliarmos no presente, estamos criando vetores que atuarão no futuro. Uma avaliação é sempre consequente; há um momento do agora, quando a avaliação é realizada, e um do depois, quando nova avaliação será feita. Precisamos pensar, simultaneamente, em duas dimensões. Um bom trabalho gera bons frutos: mas um mau trabalho também gera frutos, somente que amargos. Uma experiência de avaliação fracassada pode

determinar outros insucessos, difíceis de serem superados a curto prazo e exigindo grandes esforços para superar os problemas criados, às vezes involuntariamente, mas que refletem não apenas falta de cautela, mas também de uma reflexão crítica aprofundada. Ao avaliarmos, repetimos, é preciso que pensemos nas futuras avaliações que ocorrerão: presente e futuro estão interligados, assim como nas suas repercussões.

A elaboração de um projeto de avaliação deve ter elementos que sejam capazes de despertar o interesse da sociedade. O interesse não pode ficar limitado ao âmbito dos técnicos que planejam o trabalho. Isso seria condenar a avaliação a uma repercussão restrita; ser de conhecimento exclusivo da comunidade escolar, o que reduziria o alcance da avaliação. A participação da sociedade, por intermédio dos pais, integrantes do colegiado, no planejamento, elaboração dos instrumentos, aplicação, correção e redação dos relatórios, é uma atividade de interação social, proporcionará à sociedade uma visão dos problemas de uma instituição que é básica ao seu existir – a escola – e possibilitará uma atuação mais eficiente da educação, dando-lhe novos rumos, abrindo novas perspectivas. A avaliação tem suas implicações, inclusive na definição das políticas públicas.

A avaliação não se completa se sobre os seus dados não se fizerem pesquisas de seus diferentes aspectos. Avaliação e pesquisa se confundem em vários momentos. Há atividades que apresentam alguma transvariância, mas em outros aspectos se afastam, como a possibilidade de repetição dos resultados ou a generalização dos mesmos, que no caso da pesquisa pode ser alto, mas na avaliação é geralmente baixo. É imperioso que avaliação e pesquisa caminhem juntas, são atividades perfeitamente compatíveis, que se complementam e convergem para um ponto, que é a geração de novos conhecimentos. Sem pesquisa, a avaliação, sozinha, perde grande parte do seu impacto.

## AVALIAÇÃO E OS DADOS DA AVALIAÇÃO

A avaliação destina-se a um público e seus resultados não podem ser sonegados. O direito aos dados de uma pesquisa ou de uma avaliação pela sociedade não pode ser violado, sob qualquer pretexto. É preciso, desse modo, inicialmente, saber quais

os dados de informação que devem ser coletados, respeitados, naturalmente, os direitos humanos, o direito à privacidade. Algumas vezes, um excesso de dados é levantado sem maior sentido, porque não serão utilizados ou se o forem não contribuirão efetivamente para esclarecer ou solucionar um determinado problema. A par disso, a manipulação e análise de um número grande de informações demanda tempo e pessoal qualificado para a sua disseminação. Acreditamos, desse modo, que seja prudente, inicialmente, uma seleção do que vai ser efetivamente útil e, depois, que a divulgação dos elementos encontrados se faça paulatinamente, e não apenas em um relatório único e final. A audiência poderá, assim, participar mais ativamente de todas as fases do progresso da avaliação ou da pesquisa.

A literatura técnica sobre avaliação e, especialmente, sobre medidas, tem crescido vertiginosamente, variando bastante os seus níveis de qualidade e complexidade, problema este que precisa ser considerado ainda que não no presente. O que nos interessa, no momento, é que essa literatura destaca reiteradamente a necessidade da elaboração de um plano rigorosamente estabelecido que seria indispensável para a realização de uma boa avaliação. É evidente que o rigor científico é indispensável, a fim de que resultados válidos e fidedignos sejam recebidos com confiabilidade pela comunidade e exerçam algum tipo de impacto sobre a sociedade e o público interessado. Contudo, é necessário ressaltar que a avaliação pode ser realizada segundo diferentes formas, às vezes menos rigorosas na sua metodologia, mas nem por isso menos importantes e oferecendo resultados significativos. Pensamos, no caso, por exemplo, na observação e na utilização do método indutivo. Apenas é preciso que o avaliador (ou pesquisador) seja uma pessoa realmente capacitada e experiente.

Uma avaliação se processa em um contexto determinado, em que atuam variáveis bastante específicas, que são próprias àquela situação particular. Aqui, a atividade do avaliador difere da do pesquisador, entre outros aspectos, na impossibilidade de replicar um determinado estudo, como pode fazer um pesquisador experimental. Em avaliação, cada caso é um caso específico, uma situação própria, um fenômeno que não se repetirá ou que se repetirá sob outros aspectos com características diferentes. O avaliador, nesse sentido, é um historiador e a avaliação uma obra de história.

## AVALIAÇÃO E ANÁLISE CRÍTICA

Uma pesquisa ou uma avaliação deve ser submetida a constante análise crítica durante o seu processo (meta-avaliação formativa) assim como no final, quando apresenta o seu produto (meta-avaliação somativa). A crítica, em todas as suas fases, não apenas contribui para o aprimoramento do processo, com a correção do rumo, tendo em vista prováveis desvios, mas também para o crescimento do próprio avaliador, cuja formação se faz através de novas experiências e a vivência de novos problemas. A análise da avaliação concorre, ainda, para que o trabalho não se desvie das suas metas, dos seus propósitos e dos objetivos estabelecidos e não venha a dar informações muitas vezes descabidas, ainda que sob uma forma bastante elegante, de problemas que não foram propostos e que fogem ao mérito da avaliação. A elegância da formatação de um relatório, ainda que útil, não é importante; fundamental em uma avaliação é a precisão dos seus resultados e, especialmente, a validade das informações.

O mérito de uma avaliação não está na forma, o mesmo ocorrendo com a pesquisa. A atuação coordenada dos trabalhos – pois avaliação e pesquisa são obras de equipe e não *one man/woman show* –, a maturidade da pesquisa e do avaliador alcançadas após reiteradas experiências de campo e também o grau de intimidade do avaliador com os futuros usuários da avaliação são elementos que vão estabelecer o estilo do relatório, ou melhor dizendo, dos vários relatórios a serem apresentados, porque, conforme a audiência, um determinado tipo de relatório, com formatação e estilos próprios, se deve impor. Um aspecto precisa ser esclarecido imediatamente: uma avaliação não é feita para um grupo restrito de iniciados; uma avaliação não se destina apenas a avaliadores ou a pesquisadores, a um grupo seleto de cientistas, mas a toda a comunidade, especialmente a professores e alunos beneficiários imediatos das avaliações.

A experiência adquirida ao longo dos anos de prática gera uma sabedoria que é bem diferente daquela advinda do estudo e da reflexão sobre o conhecimento produzido e apresentado em livros e revistas. Um jovem, no ardor de seus verdes anos, dirá (ou diria): sou positivista, quantitativo, dedutivista e somativo. Uma jovem em flor, na impetuosidade da sua juventude, dirá (ou diria): sou subjetivista, qualitativa, naturalista, indutivista e

descritiva. Os jovens, na sua imaturidade, precisam demonstrar que são, dizer que são e defender suas posições, ardorosamente. A idade, a experiência e o trabalho ensinam que há vários atalhos, que diversas são as sendas, mas que somente existe um caminho e que esse caminho é o do meio, que nos leva em certos momentos a uma posição, que é ditada pela complexidade de um contexto, e em outras circunstâncias nos conduz a um posicionamento diverso, porque outro é o quadro a ser avaliado. A rigidez das posições é pouco condizente com a flexibilidade de espírito que deve caracterizar o avaliador.

## AVALIAÇÃO E VALIDADE

A questão da validade dos dados é mais importante que a fidedignidade. Se os resultados são válidos, é evidente que são precisos, ainda que a recíproca não seja verdadeira. Os dados devem permitir que se façam inferências sobre a natureza do objeto estudado e que essas inferências possam ir além dos próprios dados (validade externa). Há um acentuado desejo de muitos, entretanto, com a sofisticação dos planejamentos e, com isso, aumentar a validade interna, que às vezes é mais importante do que as generalizações.

A avaliação implica custos, muitas vezes bastante elevados e as fontes de financiamento são bastante parcimoniosas. Assim, é preciso que o avaliador exerça um controle nos gastos e desenvolva na equipe um posicionamento ético para que as verbas não se diluam inutilmente. É necessário, entretanto, atentar para a qualidade dos dados. Às vezes, o controle excessivo pode levar ao comprometimento dos dados, ou à falta de dados essenciais ao estudo; por sua vez, o controle orçamentário em um aspecto pode prejudicar outros aspectos do programa. O gerenciamento a ser exercido pelo avaliador, assim como, na verdade, todos os seus procedimentos, deve ser caracterizado por um comedimento equilibrado que não afete a harmonia de todo o programa.

Os problemas de amostragem são bastante complexos. A amostragem exige do especialista uma formação especial, um treinamento básico e um estudo especializado bastante aprofundado. A amostragem não admite espíritos aventureiros ou simplistas, que acreditam que 10% de alguma coisa seja representativo de algo que às vezes desconhecem totalmente. Se o trabalho é por

amostragem, por detrás dela deve existir uma matemática bastante sofisticada. Surpreendente, e paradoxal, é que nem sempre uma amostra, por ser supostamente representativa, proporciona todos os elementos necessários para um estudo; às vezes, a natureza do estudo pode necessitar de uma sobrerrepresentação de casos excepcionais para uma coleta de melhores informações; dar o papel do estatístico especializado em amostragem, que deve ser o associado constante do avaliador nos seus estudos, nas suas pesquisas, tendo em vista a validade dos resultados.

## AVALIAÇÃO E ÁREA DE HABILITAÇÃO

A avaliação, no nosso contexto educacional pleno de incertezas e dúvidas, um contexto bastante fragilizado, é bom insistir, não se constitui em **área de habilitação**, por decisão oficial, tomada há algum tempo por um colegiado de sábios educadores. Apenas a administração, a supervisão e a orientação educacional seriam as supostas áreas de habilitação, segundo pareceres de grande erudição livresca e pensamentos confusos. Assim, a formação do avaliador é feita, quando o é, em outros cursos – pedagogia, psicologia e, possivelmente, na sociologia. Aqui surgem alguns problemas: a pedagogia, curso de grande amplitude de áreas, forma generalistas, e a avaliação é vista *en passant*; a psicologia, com suas várias tendências teóricas, excepcionalmente dá alguns destaques à parte docimológica, e a sociologia, fragmentada em cursos monográficos, apresenta, sem grande amplitude, aspectos da metodologia da pesquisa. O quadro assim apresentado exclui praticamente a avaliação; desse modo, por desvio de formação, muitos usam delineamentos de planejamento que são apropriados às pesquisas de laboratório, mas nem sempre adequados para a avaliação educacional.

Se a ideia é mudar, através de um processo de avaliação, se, além disso, pretende-se, ainda utilizando uma avaliação, realizar a implementação de um determinado programa, precisamos, inicialmente, estabelecer o máximo possível de indicadores, o que nem sempre é realizado, sendo consideradas apenas variáveis ligadas ao produto, ou seja, ao rendimento escolar e às atitudes. Necessitamos, inicialmente, considerar um conjunto razoável de elementos da demografia educacional, a fim de caracterizar

a nossa clientela, o que se completa com variáveis socioeconômicas relacionadas à escola e uma série de indicadores não ligados à escola, mas que são de fundamental importância para a compreensão do fenômeno – por exemplo, a participação da família na educação – e, depois, então, indicadores que dizem respeito diretamente à escola, para conhecimento de variáveis do contexto, do processo e do produto. A existência de múltiplos indicadores que se intercruzam concorre, pois, para que se forme um quadro detalhado da realidade a ser criticamente analisada e avaliada.

O avaliador nem sempre está ligado a um único projeto, muitas vezes, vê-se obrigado, por razões profissionais, a atuar simultaneamente em diversas avaliações, às vezes sobre objetos inteiramente diferentes entre si. Isto exige um esforço hercúleo, uma capacidade de multiplicar-se diante de diferentes situações sem se sentir perdido, sem saber que procedimentos adotar e sem se deixar sucumbir diante do estado de entropia em que se acha submerso. A avaliação acaba sendo um mergulho em profundidade num mundo de incertezas, por falta de um treinamento adequado do avaliador.

## AVALIAÇÃO E ANÁLISE ESTATÍSTICA

Algumas avaliações realizam o levantamento de grande soma de dados que, depois, são submetidos a técnicas estatísticas de alto nível e sofisticação matemática. A interpretação e análise desses elementos, por uma questão de prudência, não devem ser da responsabilidade exclusiva de uma única pessoa, mas submetidas a diversos avaliadores experimentados e entre eles discutidas. Pensemos nas implicações de uma análise problemática realizada por uma única pessoa. É uma grande responsabilidade pois a avaliação visa a tomada de decisões, que por sua vez objetivam mudanças, que precisam estar alicerçadas em elementos sólidos, inquestionáveis, e aceitos pela sociedade. Ao relatório final deve anteceder, como uma coisa perfeitamente normal, uma análise das estatísticas por diferentes pessoas igualmente capacitadas.

A avaliação no momento presente (1997) constitui preocupação de diferentes segmentos da sociedade e não apenas da comunidade educacional. Chega a haver uma certa euforia, o que, na verdade, é preocupante, se atentarmos para o fato de que isso

pode gerar discursos entusiásticos e visões obscurecidas por uma falsa luminosidade. Avaliações, assim como pesquisas, costumam incidir sobre as mesmas áreas, gerando trabalhos paralelos. Estas atividades repetitivas somente têm sentido quando se processa um intercâmbio de informações, quando ocorre um processo de *cross-fertilization* que auxilia na solução cooperativa dos diversos problemas que estão sendo considerados. A troca de informações entre diferentes grupos, além de desejável, é sempre bem-vinda.

As avaliações costumam ter um custo elevado mesmo quando se utiliza a mão de obra à disposição nas escolas ou nas Secretarias de Estado. Apesar do emprego de mão de obra muitas vezes ociosa, o custo existe e o avaliador deve atentar para procedimentos de racionalização dos gastos. É necessário que o avaliador atente para os procedimentos de racionalização dos gastos. É importante que o avaliador pergunte a si próprio se o elemento que vai coletar justifica o investimento a ser feito e se haverá um uso consequente deste mesmo elemento, justificando-se mais ainda os gastos operados, pelas ações que pode gerar. Isso é indiscutível tanto na área pública como na atividade privada.

## AVALIAÇÃO E COMPARABILIDADE

O desenvolvimento de um programa de avaliação muitas vezes se reveste de grande complexidade, exigindo elementos técnicos em vários níveis, pessoal qualificado para diferentes tarefas e muitas vezes o envolvimento de dezenas de pessoas para que o processo ocorra com um número mínimo de problemas. Ora, certas avaliações são verdadeiras operações de estado-maior, que exigem hierarquização, disciplina e seguimento de normas prescritas. Há necessidade, assim, do estabelecimento de mecanismos de controle que garantam o desenvolvimento do processo, – por exemplo – o monitoramento dos pais, como representantes da sociedade. É preciso, contudo, que esses controles representem um custo razoável e limitado às previsões orçamentárias; por sua vez é também necessário que se pense na possibilidade da dispensa desses controles e nos efeitos negativos irreversíveis para a avaliação se a eliminação dos mesmos ocorrer.

É comum em avaliação ouvir falar em comparações dos dados. A coisa é colocada de tal forma que se não houver compa-

rações não haveria avaliação. O assunto merece consideração, evidentemente, sem radicalizações. É evidente que havendo equalização dos resultados, as medidas obtidas poderão, em princípio, ser comparadas.

É igualmente claro que existem diferentes maneiras de promover essa equalização, seja por intermédio de procedimentos clássicos, seja via procedimentos chamados modernos, ainda que cinquentenários, como é o caso da teoria do traço latente mais conhecida por teoria da resposta ao item, vista por alguns como o *dernier-cri* na área da abordagem quantitativa. É necessário, antes de mais nada, que se atente para a natureza das comparações desejadas e o estágio em que se encontra a avaliação no contexto educacional, quando, então, se poderá pensar em comparações de dados, que, às vezes, frustram as nossas expectativas. Finalmente, é preciso atentar para o fato de que programas com múltiplos e às vezes objetivos dissimilares não geram dados comparáveis, e que nenhuma tecnologia, por mais evoluída que seja, conseguirá superar o problema e tornar comparáveis dados que na realidade não o podem ser. É preciso cuidado com o mito da comparabilidade[3].

**3** Ver NUTTALL, D. The myth of comparabilily. *Journal of the National Association of Inspectors and Advisers*, n. 11, p. 16-18, 1979.

## AVALIAÇÃO E TERCEIRIZAÇÃO

Avaliação é um processo financeiramente dispendioso, exigindo investimentos, e que demanda o envolvimento de um número elevado de pessoal técnico-científico, administradores e elementos da área burocrática em diferentes níveis. Isto significa dizer que muitas instituições, por suas limitações, não fazem avaliações, ainda que as desejem, e nem as contratam de terceiros, pelas mesmas razões. A avaliação está se tornando uma atividade entre instituições especializadas e órgãos governamentais, via diferentes processos de associação, consórcio ou simples terceirização. A legislação sobre o assunto é complexa e de difícil entendimento, demandando assessoria especializada na hermenêutica jurídica. Qualquer falha que venha a ocorrer pode inviabilizar todo o processo, inclusive impedindo a realização da avaliação, em muitos casos.

As relações avaliação-avaliador são intermediadas por uma instituição, na maioria das vezes. O avaliador não tem maior

envolvimento institucional, sendo apenas um técnico, mas estranho aos quadros institucionais. O avaliador, muitas vezes, se encontra em uma situação bastante bizarra, vendo-se obrigado, por força contratual, a aceitar certas decisões ou a participar de ações que foram decididas à sua revelia, sem a sua concordância, sem ao menos ter sido ouvido. Isso é realmente um problema, que pode ter implicações éticas para o avaliador que, assim, fica numa situação conflitiva que pode levar à renúncia ao projeto, com prejuízo para ambas as partes, sendo aconselhável, portanto, que se estabeleçam negociações entre instituições e avaliadores a fim de evitar constrangimentos que possam invalidar os esforços no sentido de realizar um trabalho significativo.

A sociedade moderna, com suas relações jurídicas extremamente complexas, está criando uma nova situação e gerando novos tipos de relacionamento, forçando o aparecimento de outras formas de responsabilidade ou de corresponsabilidade. Anteriormente, a avaliação se fazia pela contratação de alguém responsável pela avaliação, um professor com experiência, e uma instituição. A responsabilidade pelo sucesso e, sobretudo, pelo fracasso estava centrada na pessoa do avaliador; hoje, entretanto, as relações se estão alterando. O avaliador, um professor ou especialista, está ligado a uma instituição, que é contratada para fazer a avaliação. Desse modo, as relações jurídicas e a responsabilidade se alteram, não é mais uma relação indivíduo/instituição, mas instituição/instituição, e a responsabilidade não é mais exclusiva do avaliador, mas da instituição, passando o avaliador a corresponsável, apenas.

## AVALIAÇÃO E RELAÇÕES CONTRATUAIS

Uma avaliação deve estabelecer necessariamente um cronograma de desenvolvimento de suas várias fases, com a especificação das tarefas, e suas datas prováveis. É evidente que não se pretende um esquema rígido, que o planejamento tenha a precisão de um maquinário suíço. Seria uma violência, inclusive contra o próprio avaliador. Os prazos, em avaliação, devem ser bastante flexíveis para não criar situações de tensão e ansiedade, que dificultam o desenrolar do processo. Os prazos, em avaliação e em pesquisa, devem ser postos com antecedência, discutidos

pela comunidade e cumpridos sem relutância. Inclusive, deve-se fixar um *deadline* para o trabalho avaliativo. A fixação de um cronograma é de um óbvio total, sendo indispensável qualquer comentário. O estabelecimento de um prazo final também se justifica, pois a sua fixação no início vai ditar o ritmo dos esforços ao longo dos trabalhos. Avaliação, planejamento e cronograma – é impossível fugir a esses dois últimos elementos, qualquer que seja o tipo de avaliação: quantitativa ou qualitativa.

As relações contratuais para a prestação de serviços de avaliação estão ficando cada vez mais complexas, mais detalhistas, mais cheias de sutilezas e, consequentemente, dando origem a contratos longos, com itens, subitens, parágrafos e alíneas, que acabam fazendo com que sua leitura seja difícil e, frequentemente, bastante tediosa. É mais uma influência do bacharelismo coimbrão na área educacional. A tentativa de amarrar bem as coisas, prevendo tudo e estabelecendo todos os detalhes, acaba por dificultar a avaliação, às vezes impedindo a de concretizar-se. Uma lição é preciso extrair desta situação: definir demais as coisas não é a melhor forma de estruturar um bom esquema para levantamento de informações, pelo menos em avaliação educacional.

O macro e o micro em avaliação precisam ser considerados. Há casos em que uma macroavaliação, realizada a partir de toda uma população, impõe-se por diferentes razões, inclusive por suas ressonâncias políticas. A microavaliação, a partir de uma amostra, também é igualmente válida, não nos esqueçamos. Isto significa que precisamos atentar para as diferentes circunstâncias que vão afetar a concretização da avaliação; no entanto, é forçoso que nos convençamos de que, pelo simples fato de estarmos realizando uma pesquisa em larga escala, envolvendo centenas de milhares de sujeitos, não estaremos coletando informações de melhor qualidade, mais válidas, do que se tivéssemos partido para uma avaliação amostral, em pequena escala. As circunstâncias, o contexto e os recursos financeiros são variáveis que ditarão o caminho a seguir.

## AVALIAÇÃO E MACROAVALIAÇÃO

As avaliações, conforme foi reiterado ao longo destas reflexões, são um empreendimento dispendioso. *Ipso facto*, uma avaliação

que abranja um número considerável de sujeitos e exija infraestrutura complexa, além, naturalmente, de um *software* sofisticado, deve ser objeto de múltiplos financiamentos oriundos de agências diferentes: públicas e privadas; nacionais e internacionais; entretanto, quando se aprofunda o problema, chega-se à conclusão de que essas agências devem ter diferentes perspectivas sociais, políticas e até mesmo econômicas. Parece-nos salutar que uma visão multifacetada venha a influir sobre a abordagem metodológica e a diversidade dos interesses a considerar.

Vivemos, pelo menos no contexto nacional, no momento presente (1997), a hora das grandes avaliações: avaliação de sistemas de ensino, avaliações ao final do 2º Grau como uma nova forma de acesso ao ensino superior e avaliação das grandes áreas profissionais, ao final dos cursos, para, a partir desses resultados, fazer uma avaliação institucional. Não pretendemos entrar no mérito dessas complexas questões. A vida é a grande mestra, mesmo em avaliação. Fazer macroavaliações implica, necessariamente, descentralizar tarefas e atividades, é uma questão de bom senso; é uma prática saudável. A avaliação, reproduzindo uma expressão utilizada anteriormente, não é de forma alguma *one man/woman show*; mas um trabalho socializado por uma sociedade que busca conhecer a si mesma, suas virtudes e seus defeitos, principalmente estes últimos.

A avaliação pode prestar grandes serviços à sociedade, especialmente em relação a elementos fundamentais: a escola, seu currículo, seus programas, seus professores e, acima de tudo, seus alunos. A sociedade começa a descobrir a avaliação, a sentir os efeitos da sua atuação e a perceber que para se autoconhecer necessita estabelecer um estreito relacionamento com a avaliação, inclusive para não sucumbir. Entretanto, até agora, a contribuição da avaliação para a sociedade ainda não é perfeitamente reconhecida. Por quê? Apesar da importância da avaliação, o papel do avaliador, como profissional da educação, também ainda não foi reconhecido pela sociedade, como ocorre, aliás, com a profissão do professor. A importância do trabalho de ambos é figura de retórica dos que detêm o poder.

## AVALIAÇÃO E *EXPERTISE*

O momento presente (1997) é de excitação com o *boom* da avaliação. A partir da cúpula da educação ao sacrificadíssimo professor de escola rural todos, repentinamente, passaram a acreditar, a desejar, a louvar e a aplaudir diferentes propostas de avaliação, nos vários níveis educacionais. A comunidade se esquece de que para uma avaliação bem feita, precisamos, antes de mais nada, de pessoal técnico competente e com *expertise* adquirida ao longo de uma prática constante. E nós perdemos o bonde da história, no campo da avaliação, porque não se procurou formar recursos humanos, jovens, sobretudo, nos grandes centros de excelência em avaliação. E existem muitos nos Estados Unidos, na Inglaterra, na Escócia e em outros países fora do mundo anglo-saxão. Por outro lado – e há sempre um outro lado –, o governo em seus diferentes níveis – federal, estadual e municipal – não se preocupou em fazer investimentos financeiros na área, salvo alguns poucos casos (INEP/MEQ). Não é possível avaliar sem recursos humanos e sem recursos financeiros, é obvio, mas esta é a realidade.

A avaliação, no momento atual, em nosso contexto, é um trabalho isolado. O avaliador é um solitário perdido na sua solidão muitas vezes do autodidatismo, e isso é um perigo. Este não é o caminho para impor-se socialmente e adquirir o *status* que deve ter na sociedade. Sozinho, nada conseguirá. Ficará perdido na estepe desolada em que se pode tornar o mundo da educação. Sem compartilhar seus interesses, suas dúvidas e, para que não dizer, as suas angústias com outros avaliadores, será um ser atomizado, em um mundo intelectual estéril. A avaliação é um trabalho participativo, com vistas à socialização das experiências individuais.

O ritual de pesquisa educacional estabelece que, identificado um problema, se procure fazer um levantamento de trabalhos anteriores para depois mergulhar efetivamente no campo da investigação. É preciso conhecer e assimilar experiências anteriores, acumulando-as e incorporando-as ao patrimônio do pesquisador. A mesma liturgia deve ser seguida pelo avaliador. Ainda que o número de avaliações no nosso contexto seja limitado, a literatura estrangeira, e não apenas a anglo-saxã, é copiosa, sendo necessário um trabalho seletivo, tendo em vista que, ao lado de trabalhos de relevância, existem experiências sem grande significação, que pouco, ou mesmo nada, acrescentam

ao existir do avaliador. Todo o material acumulado deve ser estudado, discutido e analisado em profundidade, a fim de que, mais tarde, o avaliador realize trabalhos significativos que, por diferentes caminhos, possa influenciar na educação, inclusive por intermédio da definição de novas políticas públicas.

## AVALIAÇÃO E COMPONENTES ÉTICOS

A avaliação, assim como a pesquisa, tem um conjunto de componentes éticos que não podem ser ignorados. A pesquisa, talvez tendo em vista o fato de possuir uma tradição mais antiga, já mereceu a consideração de vários teóricos e praticantes, que se debruçaram sobre o problema e estabeleceram padrões de orientação. Poderíamos simplesmente aceitar essas normas de conduta; entretanto, acreditamos que os avaliadores também se devam dedicar, talvez no seu lazer criativo, à análise da avaliação sob o ponto de vista da ética, especialmente porque, no momento, com o *boom* de avaliações no contexto nacional, começa a aparecer uma figura exótica, um estranho no ninho, a que chamaremos, eufemisticamente, de o *parvenu* da avaliação, cujo grau de probidade científica somente será possível constatar na medida em que os avaliadores promovam uma revisão crítica dos seus desempenhos, do desenho metodológico de seus projetos e da conduta ética subjacente a toda avaliação.

Uma pergunta surge de imediato e pode constituir-se em motivo de angústia: como proceder, como agir para que padrões éticos sejam seguidos? Parece-nos que os avaliadores devam estar sempre em vigília na preservação de padrões profissionais e éticos elevados, considerando que o avaliador é um pesquisador, é especialmente um educador preocupado em influenciar e gerar novas políticas públicas, que se transformarão em ações, e, ao final, influirão em um universo de pessoas, cujos direitos precisam ser respeitados. Por isso, a avaliação deve promover um amplo debate sobre suas estratégias de ação, suas controvérsias, suas motivações e suas consequências sobre a audiência a que ela se destina. Diríamos, então, que a avaliação precisa sofrer constantes e múltiplas críticas independentes, para que se possa desenvolver e cumprir sua destinação.

# INSTRUÇÕES A COLABORADORES
**(impresso)**

## NORMAS GERAIS

*Estudos em Avaliação Educacional* publica trabalhos inéditos referentes à educação, apresentados sob a forma de relatos de pesquisa, ensaios teóricos, metodologias, revisões críticas, artigos e resenhas.

Excepcionalmente, serão aceitos trabalhos que tenham sido publicados em periódicos estrangeiros (com a indicação da fonte), os quais serão submetidos à mesma avaliação dos artigos inéditos. O autor deverá apresentar a autorização da revista em que seu artigo tenha sido originalmente publicado.

Os originais recebidos são apreciados por especialistas da área e pelo Comitê Editorial, mantendo-se em sigilo a autoria dos textos.

Os autores recebem comunicação relativa aos pareceres emitidos. O Comitê Editorial reserva-se o direito de recusar o artigo ao qual foram solicitadas ressalvas, caso não sejam atendidas a contento.

Se a matéria for aceita para publicação, a revista permite-se introduzir pequenas alterações formais no texto, respeitando o estilo e a opinião dos autores.

Os trabalhos não deverão ser publicados em qualquer outra forma antes de decorridos seis meses de sua publicação em *Estudos em Avaliação Educacional*.

Artigos de um mesmo autor só será publicado com intervalo de, pelo menos, seis meses.

Solicita-se do(s) autor(es): nome completo, vínculo institucional: última ocupação profissional: cargo e filiação (empresa, instituição ou organização); unidade de referência (da ocupação profissional): Faculdade/Instituto, Programa de Graduação/Pós-graduação, Departamento; titulação (graduado, especialista, mestre/mestrando, doutor/doutorando), endereço, telefone, celular e correio eletrônico. Pede-se, ainda, que o autor indique como seu nome deve constar da publicação, bem como o nome completo da instituição à qual está vinculado e seu e-mail de contato.

Os autores receberão três exemplares impressos da revista em que seus textos forem publicados.

## APRESENTAÇÃO DOS ORIGINAIS

Para submeter um artigo a *Estudos em Avaliação Educacional*, é necessário:

- Estar cadastrado no Portal de Periódicos da Fundação Carlos Chagas.. Se não estiver, acesse para fazer o cadastro: <http://publicacoes.fcc.org.br/ojs/> (o login e a senha serão a chave para o acesso ao sistema).

- Encaminhar duas versões do artigo: uma para avaliação, em PDF, sem informações que permitam identificar a autoria; e outra, em Microsoft Word, com todas as informações.

Caso o artigo seja em coautoria, o ideal é que todos os autores estejam cadastrados no sistema. É possível, no entanto, o envio do texto apenas com o cadastro de um dos autores. Em ambos os casos, a pessoa que envia o arquivo precisa incluir os coautores no Passo 3. Metadados da Submissão → Incluir Autor.

Para acompanhar o *status* da submissão, deve-se acessar o sistema → menu → acesso → login e senha.

A primeira página do texto deve trazer o título do trabalho e omitir o nome do autor e a filiação institucional, a fim de assegurar o anonimato no processo de avaliação.

Na **extensão**, os artigos não podem exceder 25 páginas (incluídos os anexos) e a extensão máxima das resenhas é de seis páginas, e devem ter o seguinte formato obrigatório: 3 cm de margem superior, 3 cm de margem inferior, 3 cm de margem esquerda e 2 cm de margem direita; parágrafo 1,25; com espaçamento de 1,5 entre as linhas; sem espaço (anterior ou posterior) entre os parágrafos, páginas enumeradas (após a folha de rosto, na margem inferior à direita), fonte em *Times New Roman*, no corpo 12.

**Títulos e subtítulos** devem ter, no máximo, 11 palavras (incluindo artigos, preposições, conjunções etc.).

Toda matéria, à exceção de resenhas, precisa vir acompanhada de **resumo** contendo no máximo 11 linhas, com espaçamento simples entre as linhas e sem espaço entre parágrafos, sem conter siglas nem referências, trazendo, em seu início, o título do trabalho. Ao final do resumo, indicar quatro palavras-chave (descritores) do conteúdo do texto.

**Citações, remissões, notas e siglas** devem obedecer às regras da ABNT (NBR 10520, 2002). As **citações** diretas (textuais), com até três linhas, devem ser incorporadas ao texto, entre aspas, sendo necessário indicar o sobrenome do autor, ano e número da página.

Ex.: Em função desses indicadores, "chegou-se à organização de cinco grandes grupos de escolas denominadas azul, verde, amarelo, laranja e vermelho" (SÃO PAULO, 2001, p. 55).

**Citações** com mais de três linhas deverão ir em um bloco abaixo do texto, sem aspas, com recuo de 1,25 cm, a partir da margem esquerda, com espaçamento simples entre as linhas e sem espaço entre parágrafos, em fonte *Times New Roman* e corpo 10, sendo necessário indicar o sobrenome do autor, ano e página. Ex.:

> [...] a sofisticação técnica da avaliação nacional – que hoje ocupa a atenção da cúpula decisória e de seus assessores – apresenta-se como entrave para a compreensão; tanto pelos atores dos sistemas e escolas como pela população em geral, do processo avaliativo realizado. (FREITAS, 2004, p. 685)

Na **citação de citação** deve ser empregada a expressão latina "apud" (citado por) para identificar a fonte que foi efetivamente consultada, a qual deve ter a referência completa no rodapé; e na lista de Referências incluir apenas a obra consultada (CARONE et al., 2003).

Ex.: Para Watson (apud CARONE et al., 2003) [...].

As **remissões bibliográficas** indiretas são incorporadas ao texto entre parênteses (ano).

Ex.: Segundo João Barroso (2006), todos...

As **notas explicativas** devem ser evitadas e utilizadas apenas quando for estritamente necessário, preferencialmente sem ultrapassar três linhas. Devem figurar sempre no rodapé da página, numeradas sequencialmente.

As **siglas** devem ser desdobradas quando mencionadas à primeira vez no artigo.

Ex.: Exame Nacional do Ensino Médio (Enem).

**Tabelas, gráficos, quadros e figuras** (assim como os **títulos** e as **fontes**) devem ser apresentados no corpo do texto, e não em caixas de texto, alinhados à esquerda, em sua página correspondente, numerados com algarismos arábicos, com títulos (posicionados acima, em corpo 12) padronizados quanto ao formato e termos utilizados. Abaixo destes, sem estarem em caixas de texto, deve, obrigatoriamente, ser indicada a fonte dos dados (remetida às referências bibliográficas), com autoria e ano, inclusive se for de elaboração própria dos autores, em corpo 10, alinhada à esquerda, espaço 1,5 entre linhas.

Tabelas, gráficos e quadros devem, ainda, ser enviados em um arquivo separado, em *software* compatível com o ambiente Windows, de preferência em Excel, e as figuras (ilustrações, imagens, mapas, fotos etc.) em arquivo com alta resolução (300 dpi), todos produzidos em preto e branco, em tamanho máximo de 10 cm de largura.

**Referências de cunho bibliográfico** devem vir ao final do texto, por ordem alfabética de sobrenome do autor, e, quando possível, fazer constar por extenso o prenome dos autores. Os títulos das obras devem vir em itálico.

Quando houver dois ou três autores, separa-se o primeiro autor e os demais por ponto e vírgula; ultrapassando três autores, faz-se a entrada pelo autor principal (referenciado no texto) e substitui-se os outros pela expressão "et al.".

Sua apresentação deve seguir as normas da ABNT (NBR 6023, 2002). A exatidão das referências e a correta citação no texto são de responsabilidade do(s) autor(es) dos artigos, sendo uma exigência para a publicação do trabalho.

DIREITO DE RESPOSTA

*Estudos em Avaliação Educacional* acolhe comentário(s) a artigo publicado na revista. Se o comentário for aceito para publicação, a revista oferecerá ao autor igual espaço para réplica, que poderá ser publicada no mesmo número do comentário ou no número subsequente. Ambos estão sujeitos ao mesmo processo de avaliação dos demais textos. Não são aceitos comentários ou réplicas a resenhas.