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.