AVALLAÇÃO EM LARGA ESCALA: uma proposta inovadora

Ruben Klein* Nilma
Santos Fontanive**

Introdução

A avaliação educacional é um sistema de informações que tem como objetivos fornecer diagnóstico e subsídios para a implementação ou manutenção de políticas educacionais. Ela deve ser concebida também para prover um contínuo monitoramento do sistema educacional com vistas a detectar os efeitos positivos ou negativos de políticas adotadas.

Um sistema de avaliação deve obter e organizar informações periódicas e comparáveis sobre os diferentes aspectos do sistema educacional. Neste trabalho, porém, os autores restringir-se-ão a abordar a avaliação educacional como um sistema de informação sobre os alunos em dois principais aspectos: movimentação e fluxo escolar e aprendizagem.

As estatísticas educacionais sobre movimentação e fluxo escolar foram tratadas erradamente por muito tempo no Brasil. Trabalhos recentes de Fletcher e Ribeiro (1987 e 1988), Ribeiro (1991), Klein e Ribeiro (1991) e Klein (1995) apontam os erros e apresentam duas metodologias de correção destas estatísticas.

Estes autores mostram que o acesso à primeira série do primeiro grau está praticamente universalizado, uma vez que pelo menos 95% de uma coorte de idade têm acesso a esta série. Já a conclusão do primeiro grau está longe de ser universal, pois em 1990 somente 45% de uma coorte estava concluindo o primeiro grau, seja via sistema regular, seja via supletivo de ensino. Ao mesmo tempo, os trabalhos mostram que o número de matrículas no primeiro grau era maior do que o número de crianças de 7 a 14 anos.

Os autores demonstram que o grande problema do sistema educacional brasileiro é a repetência e não a evasão. Cerca de 50% dos alunos matriculados no sistema regular de ensino repetem a primeira série a cada ano, enquanto somente 2% se evadem. Considerando-se as oito séries do primeiro grau, 33% dos alunos repetem uma série a cada ano, enquanto somente cerca de 5% saem do sistema regular de ensino sem concluí-lo. Os alunos passam em média cerca de nove anos no primeiro grau e os que concluem o fazem em média em 11 anos. A grande maioria dos alunos tem pelo menos uma repetência no primeiro grau, mas insiste em ficar na escola, só saindo após vários anos, por não conseguir progredir.

Os dados da Pesquisa Nacional de Amostras por Domicílio (PNAD) do IBGE, coletados anualmente entre setembro e novembro, indicam que, em 1990,90% das crianças de 9 e 10 anos estavam freqüentando a escola e que 72% das crianças de 14 anos estavam

* Pesquisador do Laboratório Nacional de Computação Científica (LNCC/CNPq) e consultor da Fundação Cesgranrio.
** Professora adjunta da UFRJ e consultora da Fundação Cesgranrio.

matriculadas em qualquer uma das séries do primeiro grau do ensino regular.

Estes indicadores de movimentação e fluxo escolar, embora úteis para nos dar uma idéia da eficiência do sistema, não nos fornecem informações sobre a qualidade do ensino oferecido aos alunos. Poderíamos especular que os altos índices de repetência fossem devidos a um alto grau de exigência para aprovação, e que os concluintes fossem alunos altamente qualificados. Infelizmente, dados sobre o desempenho de alunos em exames vestibulares e nas avaliações realizadas pelo Sistema de Avaliação da Educação Básica (SAEB), em 1990 e 1993, indicam que as altas taxas de repetência são acompanhadas de um ensino de baixa qualidade.

Neste contexto, toma-se indispensável a criação e manutenção de um sistema de avaliação de aprendizagem capaz de fornecer informações consistentes, periódicas e comparáveis sobre o desempenho dos alunos.

Avaliação em **larga escala: objetivos e metodologia**

Os objetivos da avaliação em larga escala do sistema escolar, aqui propostos, são os de informar o que populações e subpopulações de alunos em diferentes séries sabem e são capazes de fazer, em um determinado momento, e acompanhar sua evolução ao longo dos anos. Não é seu objetivo fornecer informações sobre alunos ou escolas individuais.

Para isso, é necessário que um grande número de itens (mais de 100) de uma área curricular de interesse seja aplicado à população de alunos em consideração, de modo que haja uma boa cobertura dos tópicos de programa de ensino.

Se um aluno fosse responder a todos os itens, ele levaria várias horas, o que não é desejável, pois sua participação é "voluntária". Não recebendo nota e não tendo sua aprovação afetada pelo resultado do teste, o aluno está em uma situação muito diferente de quando ele se candidata a um exame vestibular ou a um concurso.

A avaliação de todos os alunos de uma população apresenta também diversas restrições de ordem operacional e de custos. No entanto, na avaliação em larga escala aqui proposta, trabalha-se com uma amostra representativa da população de alunos considerada, e com uma amostragem matricial dos itens, de maneira que cada aluno responda somente a uma parte dos itens.

Uma das técnicas empregadas, hoje em dia, é o planejamento em blocos incompletos balanceados, na qual os itens são agrupados em blocos. Feito isso, são compostos cadernos de teste dep blocos, de tal modo que cada bloco apareça o mesmo número de vezes em cada posição dos cadernos, e cada par de blocos apareça uma vez e somente uma vez em um dos cadernos. Este planejamento, além de testar se a posição do bloco tem influência nas respostas dos alunos, permite calcular a correlação entre dois itens de teste quaisquer. Distribuídos em espiral, os cadernos garantem que a aplicação seja aleatória e que alunos na mesma turma respondam, em geral, a cadernos diferentes, ainda que cada caderno de teste seja aplicado aproximadamente no mesmo número de alunos na amostra.

Dado que os alunos respondem a vários cadernos de teste diferentes, não faz sentido apresentar escores de resultados individuais dos alunos, pois eles nâo são comparáveis. Pode-se, por exemplo, utilizar indicadores usuais, tais como os percentuais de acerto por item, para a população e subpopulações pesquisadas. No entanto, fica muito difícil apresentar resultados compreensíveis sem a utilização de técnicas estatísticas descritivas de resumo de informações. Embora a média dos percentuais de acerto dos itens possa ser utilizada, ela apresenta vários inconvenientes, como limitar a comparação a grupos de itens comuns na mesma avaliação ou em várias avaliações realizadas ao longo dos anos. Por exemplo, fica muito difícil comparar resultados de alunos de séries diferentes, a não ser para os itens comuns respondidos por eles.

Outras limitações são a dificuldade de interpretação dessa média de percentuais de acerto e a falta de informação sobre a distribuição de habilidades entre os alunos na população ou subpopulação, quando estes respondem somente a uma parte dos itens.

Hoje em dia, técnicas de obtenção de escalas, baseadas nas respostas aos itens como, por exemplo, a Teoria da Resposta ao Item (TRI), permitem superar as limitações expressas acima. Todos os alunos podem ser colocados em uma escala comum, mesmo que nenhum dos alunos responda a todos os itens. Usando a escala comum, é possível estimar distribuições de proficiência para a população e subpopulações e compará-las. É possível também estimar relações entre as proficiências, as variáveis socioeconômicas e culturais e as do ambiente escolar pesquisadas.

A TRI supõe que o desempenho do aluno em um teste pode ser explicado por características ou variáveis latentes subjacentes (nâo observáveis diretamente) do aluno. Estas variáveis são chamadas de proficiências ou habilidades. Em geral, procura-se reunir itens para os quais se supõe que uma certa proficiência ou habilidade é dominante. Por exemplo, podemos considerar uma proficiência para geometria, outra para número e operações, etc.

A TRI é um conjunto de modelos onde a probabilidade de resposta a um item é modelada como função da proficiência do aluno (variável não observável) e de parâmetros (que expressam certas propriedades) do item. Quanto maior a proficiência, maior a probabilidade de o aluno acertar o item.

Os itens podem ser do tipo binário, certo ou errado, como, por exemplo, na múltipla escolha, ou do tipo polítomo, como em questões onde o aluno tem que escrever a resposta e esta é classificada em uma de várias categorias ordenadas que variam de errado a correto.

A proficiência de um aluno depende de suas características individuais, como, por exemplo, seu nível socioeconômico e cultural, sua escola, seu professor, sua série, etc. Entretanto, dada sua proficiência, a probabilidade de o aluno dar a resposta correta ao item depende somente da proficiência e não mais das outras variáveis. É claro que nem todo item satisfaz esta hipótese, razão pela qual esta deve ser verificada para cada item.

Uma propriedade importante da TRI é a de invariância dos parâmetros, isto é, os parâmetros dos itens obtidos de grupos diferentes de alunos testados e os parâmetros de proficiência baseados em grupos diferentes de itens são invariantes, exceto pela escolha de origem e escala.

Graças a essas propriedades, a TRI permite comparar alunos, mesmo que eles tenham respondido a itens diferentes, em momentos diferentes.

Para estimar a distribuição de proficiência de uma população, precisaríamos estimar as proficiências de cada aluno testado. Em um teste onde o número de itens é grande isto não é problema, porém na avaliação de larga escala proposta aqui, em que o aluno só responde a alguns itens de uma área curricular, como geometria, a incerteza na estimação da proficiência não pode ser ignorada. Assim, para se estimar as distribuições de proficiência da população e subpopulações, usa-se a metodologia do valor plausível descrita em Mislevy, Johnson e Muraki (1992).

Comparações baseadas nas distribuições de proficiência podem ser feitas, através das médias e desvios padrões das distribuições de cada grupo, e das proporções de alunos em cada grupo da população acima de certos níveis de escala, da localização dos percentis da distribuição de proficiência dentro de cada grupo.

Um dos resultados mais importantes é a interpretação da escala em certos níveis prefixados, feita, por exemplo, através de itens âncora. Fixado um nível, a idéia é selecionar itens cujo poder de discriminação se situa ao redor deste nível e usar estes itens para descrever o que os alunos cujas proficiências estão perto deste nível sabem e são capazes de fazer (Beaton e Allen, 1992).

Uma experiência de avaliação em larga escala no município do Rio de Janeiro

Os objetivos e a metodologia descritos anteriormente estão sendo empregados em uma pesquisa de avaliação de alunos de 8ª série e 3ª série do 2ª grau, do município do Rio de Janeiro, em 1995, cujo projeto é desenvolvido na Fundação Cesgranrio com o apoio da Fundação Ford. Foram selecionadas duas áreas curriculares — Matemática e Língua Portuguesa (leitura) e duas séries finais de ciclo—8ª do 1º grau e 3ª série do 2º grau — para serem avaliadas. Estas séries foram escolhidas pela possibilidade de se obter um relativo consenso sobre o que o aluno sabe e é capaz de fazer no final destes dois ciclos de estudos e ainda pela facilidade de inclusão de conteúdos e habilidades de séries anteriores, permitindo a obtenção de escalas de proficiências comuns a estes alunos. Com esta intenção, itens de teste foram aplicados simultaneamente em ambas as séries. No próximo ano, a pesquisa abrangerá a 4ª série, nas mesmas áreas curriculares, na tentativa de se avaliar as três etapas importantes do processo educacional.

Os itens de teste foram elaborados a partir de uma matriz de especificação de conteúdos curriculares e habilidades cognitivas a serem avaliados.

Um dos requisitos da TRI é a definição da habilidade cognitiva que o item mede, devendo ele, em princípio, medir uma habilidade de cada vez. Para tal, é necessário especificar as habilidades desejadas e elaborar questões que avaliem estas habilidades. Em geral, o processo de planejamento dos testes combina os conteúdos curriculares e as habilidades hierarquizadas em níveis de complexidade a partir do que se espera que o aluno saiba e seja capaz de

fazer, em uma matriz de especificação contendo dois eixos. A decisão sobre o número de itens de teste de cada célula depende, em princípio, da ênfase com que certos conteúdos são tratados e do equilíbrio entre o nível de complexidade das habilidades e a maturidade intelectual (ou escolaridade) da população de alunos a ser testada.

Esta etapa de planejamento dos testes é crítica em avaliações em larga escala, nas quais a variabilidade dos métodos e processos de ensino e as diferentes ênfases curriculares não podem ser contempladas. Procurou-se então recolher currículos e programas de ensino divulgados, livros de textos e outros materiais de ensino e, ainda, estudos e ensaios realizados por especialistas. Esta análise originou uma relação preliminar de conteúdos e habilidades que seriam avaliados. Não se pode perder de vista que, ainda que a avaliação deva refletir o que é ensinado nas escolas, ela deve também indicar caminhos de renovação da prática escolar, não se restringindo, portanto, ao que se costuma definir como "currículo mínimo". Assim, no planejamento dos testes de habilidade de leitura para alunos de 8ª série do lº e 3ª série do 2º grau, foram considerados três propósitos da leitura: leitura como experiência literária, leitura para se obter informações e leitura para a execução de tarefas. Os dois primeiros propósitos são bastante freqüentes no cotidiano das escolas, enquanto o terceiro surgiu como necessidade de se avaliar uma série de atividades de leitura desempenhadas por um cidadão no seu dia-a-dia. Atividades tais como: ler para preencher um cheque ou guias de depósito bancário; ler manuais de instrução para fazer funcionar aparelhos e eletrodomésticos; ler horários de ônibus, trens, aviões ou, ainda, ler trajetos, guias de ruas, gráficos e tabelas são exemplos deste terceiro tipo de leitura, e foram incluídos nos testes de avaliação da habilidade de leitura.

Foram elaborados cerca de 1.200 itens (300 por série e disciplina) e testados 1.038, sendo 547 itens de múltipla escolha e 491 de resposta construída pelo aluno (questão aberta).

Os itens, por série e disciplina, foram reunidos em 19 blocos, cada um com cerca de 13 a 15 itens. Foram confeccionados 57 cadernos de teste, contendo três blocos cada um, segundo o planejamento de blocos incompletos balanceados.

O perfil sociocultural dos alunos e professores, as características da prática docente, da metodologia de ensino e da gestão escolar são parte integrante do banco de itens. Estes dados foram coletados na pesquisa de campo através de cinco instrumentos: dois questionários — um sociocultural e outro de hábitos de estudo de Língua Portuguesa e Matemática — dirigidos aos alunos, um questionário encaminhado aos professores e um questionário para os diretores das escolas integrantes da amostra.

As informações obtidas serão cruzadas com os resultados do desempenho dos alunos nos testes, na tentativa de identificar fatores explicativos destes desempenhos. Embora não se pretenda estabelecer relações de causa e efeito, os resultados destes cruzamentos podem sugerir formas de intervenção na prática escolar.

Os cadernos de teste e os questionários foram aplicados em uma amostra estratificada de alunos que considerou critérios geográficos, administrativos e, no caso da 3ª série do 2º grau, também o turno.

A amostra final constituiu-se de 132 escolas com 232 turmas sorteadas e 9.250 alunos, dos quais 6.854 estiveram presentes às sessões de teste. Os testes aplicados, por pessoas selecionadas e treinadas pela equipe do projeto, foram realizados preferencialmente em dias e horários das disciplinas nas turmas pesquisadas. Com esta medida, garantiu-se o ótimo percentual de 84% de respostas dos professores aos questionários.

A colaboração de alunos, professores e de diretores das escolas foi obtida através de várias atividades de divulgação dos projetos, entre elas, seminários, *folders,* cartazes e sorteios de prêmios.

As análises dos dados foram iniciadas em setembro de 1995 e espera-se divulgar as escalas de proficiências obtidas no mês de julho de 1996, apresentando-as em relatórios técnicos e relatórios simplificados para os pais e a sociedade em geral.

Referências bibliográficas

BEATON, A.E., ALLEN, N.L. Interpreting scales through scale anchoring. *Journal of Educational Statistics,* v.17, p.191 -204, 1992.

FLETCHER, Philip R., RIBEIRO, Sergio Costa. O ensino de lº grau no Brasil de hoje. *Em Aberto,* Brasília, v.6, n.33, p.1-10, jan./mar. 1987.

_____ . *Projeto Fluxo dos Alunos de Primeiro Grau — PROFLUXO:* versão preliminar. [S.L], **1988.** mimeo.

KLEIN, Ruben. *Produção e utilização de indicadores educacionais.2.*versão preliminar. [S.L], 1995. mimeo.

KLEIN, Ruben, RIBEIRO, Sergio Costa. O censo educacional e o modelo de fluxo: o problema da repetência. *Revista Brasileira de Estatística,* Rio de Janeiro, v.52, p.5-45, 1991.

MISLEVY, R.J., JOHNSON, E.G., MURAKI, E. Scaling procedures in NAEP. *Journal of Educational Statistics,* v.17, p.131-154, 1992.

RIBEIRO, Sergio Costa. A pedagogia da repetência. *Estu-. dos Avançados,* São Paulo, v.12, n.5, p.7-21, 1991.