AVALIAÇÃO NO ENSINO DE CIÊNCIAS E O MÉTODO DOS "PROJETOS" 13

Nelio Bizzo

LICENCIATURA EM CIÊNCIAS · USP/ UNIVESP

13.1 Introdução

Nesta aula, focalizaremos a avaliação, procurando dar um panorama geral das possibilidades existentes e das práticas implementadas em nosso país, desde a intimidade da sala de aula até as avaliações em larga escala. Será discutida a proximidade das estratégias de avaliação adotadas em função do modelo pedagógico de referência e a necessidade de buscar formas inovadoras de avaliação para experiências pedagógicas que se valem de arquiteturas curriculares igualmente inovadoras.

13.2 A Natureza da Avaliação na Escola e Pedagogias

As atividades de avaliação no ambiente escolar são tradicionalmente revestidas de uma aura negativa, como parte de um ritual que teria como grande objetivo vigiar e punir os estudantes. Essa conotação da avaliação não pode ser separada das práticas pedagógicas nas quais está imbricada.

Segundo um conhecido educador brasileiro, as práticas pedagógicas em nosso país estão ligadas a três tradições hegemônicas. Uma delas, que é por ele denominada **tradicional**, caracteriza-se pela crença na eficiência da transmissão de conteúdos, que localiza o centro de gravidade das atividades de ensino no professor e determina ao aluno uma posição passiva, de **receptor** de informação. A segunda tradição ele denomina **escolanovista** ou **pedagogia nova**, que se caracteriza por dar relevo aos sentimentos e individualidades dos alunos, deslocando o centro de gravidade das atividades de ensino para o aluno, que se torna a principal referência do ato educativo. A terceira é por ele chamada **tecnicista**, que valoriza princípios de racionalidade, eficiência e produtividade, deslocando o centro de gravidade para o objeto de aprendizagem, os saberes escolares (Saviani, 2008). Para o mesmo autor, haveria ainda uma tradição **contra-hegemônica**, representada pelas propostas do educador brasileiro Paulo Freire (1921-1997), que propõe uma educação como prática emancipadora, contrapondo-se a práticas domesticadoras. Assim, dentro desse quadro de hegemonia versus contra-hegemonia, seria necessário contrastar duas formas de avaliação na escola. A **verificação da aprendizagem** seria uma amostragem descomprometida daquilo que os estudantes conseguem apresentar como resultado das atividades de ensino. Essa avaliação legitimaria a exclusão, o ranqueamento e a classificação dos

estudantes. A outra seria a **avaliação** propriamente dita, na qual o avaliador se compromete com o que está sendo avaliado, e se alinharia com a perspectiva contra-hegemônica, colocando-se ao lado do educando (Luckesi, 2011).

Assim, mesmo sem pretender adotar uma terminologia muito particular, pode-se perceber que há uma visão da avaliação como uma amostragem descomprometida da aprendizagem, que apostaria no fracasso do educando e, de outro lado, outra visão da avaliação como algo que pode ser realizado tendo em vista alcançar objetivos educacionais comprometidos com o sucesso do educando. Mesmo sendo uma concepção um pouco maniqueísta, na qual o bem se contrapõe ao mal, há de se admitir que o pensamento educacional brasileiro pode conceber a avaliação como atividade positiva quando ela se compromete com o sucesso do educando, mesmo que para isso deva renunciar a pretensões ranqueadoras e classificatórias.

13.3 Avaliação: uma tipologia consensual

Os educadores reconhecem que as atividades de avaliação obedecem a diferentes finalidades e podem se organizar de diferentes formas, conforme a perspectiva que adotem.

Uma perspectiva refere-se à posição da figura do avaliador em relação àquilo que se quer avaliar. Nesse sentido, ela é dita **avaliação interna** se um ou mais dos próprios envolvidos na atividade a ser avaliada participam como árbitros. Esse caso é indicado para situações em que um grupo busca aperfeiçoar processos e identificar as melhores práticas por meio de resultados iniciais ou mesmo de seus indícios. A dita **avaliação externa** ocorre quando um ou mais avaliadores são designados justamente por não pertencerem ao grupo envolvido na atividade a ser avaliada. Esse tipo de avaliação confere legitimidade ao resultado final e tem sido utilizado em rituais institucionais, nos quais se pretende aquilatar em que medida uma instituição alcança as metas predefinidas em sua missão.

Uma outra perspectiva refere-se à regularidade, pois a avaliação pode ser de processo ou se referir a determinado momento, sendo assim denominada **contínua** ou **pontual**. A avaliação contínua é especialmente indicada para grupos de alunos que estão trabalhando de maneira colaborativa, realizando diferentes tipos de tarefas, enquanto a avaliação pontual acaba por priorizar algum parâmetro particular. Ela é indicada em situações nas quais existe pequena interação entre avaliador e avaliado e este precisa de algum elemento que possa servir de indício da eficiência do tipo de trabalho pedagógico desenvolvido.

Outra perspectiva refere-se à escala adotada, sendo que a avaliação pode ser **normativa** ou **criterial**. Neste último caso, tem-se já conhecida uma escala de possibilidades, tidas como indicadores de mérito a alcançar. Essa escala pode ser negociada com o grupo ao longo da atividade educativa e, sendo parte de um acordo, se legitima ante o grupo. De qualquer forma, uma vez definida e conhecida pelo grupo, a avaliação aquilatará a hierarquia máxima alcançada por seus membros individualmente, ou seja, será uma medida do sucesso de cada aluno. A avaliação normativa, por outro lado, parte do estabelecimento de padrões conhecidos, explícita ou implicitamente, e compartilhados pelos membros de uma comunidade de aprendizes. Ela se propõe a verificar o desempenho relativo de cada membro do grupo, de acordo com uma distribuição normal. Assim, o desempenho médio do grupo passa a ser um critério objetivo, e cada membro é classificado de acordo com ela, podendo posicionar-se em termos relativos. Assim, a classificação é um dos objetivos primordiais da avaliação normativa, enquanto a avaliação criterial tem por objetivo primordial verificar em que medida os objetivos predefinidos foram efetivamente alcançados. A figura 1 traz dois gráficos que exemplificam os resultados esperados da avaliação normativa (esquerda) e criterial (direita) em situações bem-sucedidas de ensino.

Na perspectiva da aprendizagem, a avaliação pode ser **diagnóstica**, **somativa** ou **formativa**. No primeiro caso, ela se dispõe a recolher um conjunto de indicadores que possam criar uma base para aclarar uma dada situação inicial. Como regra, a avaliação diagnóstica tende a ser pontual, uma vez que é amostral e rápida, tendo como um de seus objetivos embasar a tomada de decisões em relação ao planejamento de ações educacionais, em especial com grupos pouco conhecidos, cuja possível homogeneidade não é plenamente conhecida.

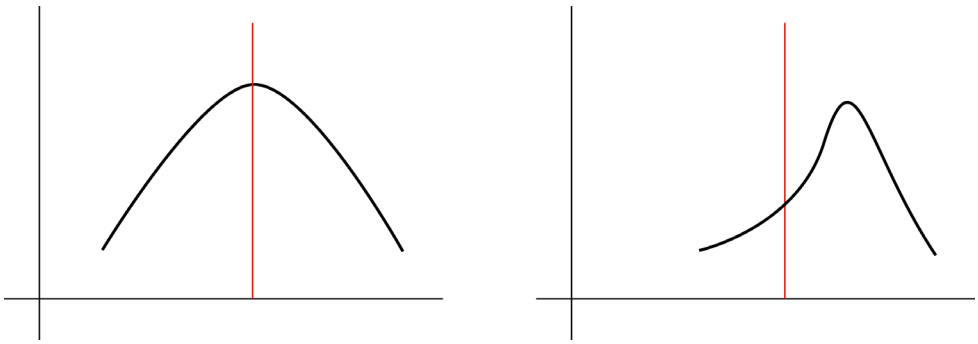

Figura 13.1: Na avaliação normativa (esquerda) espera-se que as notas dos alunos siga uma distribuição normal, com metade dos alunos acima da média, indicada pela linha vermelha, que pode não ser, necessariamente, o mínimo a ser alcançado pelo grupo para fins de aprovação, por exemplo. Na avaliação criterial, espera-se que a curva esteja fortemente deslocada para o lado do maior sucesso dos alunos, acima do limite definido previamente como mínimo (linha vermelha).

Da mesma forma, a avaliação somativa tende a ser pontual, pois pretende recolher um conjunto mais extenso e cumulativo de evidências do grau de sucesso dos estudantes em uma etapa ou conjunto de etapas de atividades educativas. Na avaliação formativa, pretende-se também recolher evidências sobre o desenvolvimento do processo ensino-aprendizagem, mas entendendo-o como processo contínuo, que inclui os próprios rituais de avaliação, oferecendo a oportunidade de o aluno consolidar suas conquistas no processo. Esse tipo de avaliação não tem uma finalidade classificatória e tende a oferecer oportunidades diversificadas para que os alunos demonstrem o resultado de seu aprendizado em contextos específicos. Entre as suas principais funções estão a de inventariar conquistas, harmonizar o aprendizado, tranquilizar, apoiar e orientar o estudante e sua família em relação a atingir os objetivos esperados, além de eventualmente corrigir concepções inadequadas que tenham vindo a desenvolver inadvertidamente.

13.4 Tradições pedagógicas, arquiteturas curriculares e avaliação

As tradições pedagógicas elegem formas mais convenientes ou adequadas de avaliação. Assim, a orientação pedagógica tradicional e a tecnicista têm nítida preferência pela avaliação pontual, somativa e normativa, uma vez que elas permitem classificar e ranquear com facilidade, e excluir, por consequência. Já a tradição escolanovista tem na avaliação contínua um de seus mais conhecidos baluartes, buscando valorizar a dimensão pessoal e processual do aprendizado.

Na perspectiva **contra-hegemônica** de Paulo Freire, a ideia de avaliação tem uma conotação muito diversa, uma vez que ela deve harmonizar-se com a perspectiva emancipadora, que tende a ser **desviante**, contrariando expectativas. Assim, ela tende a se opor ao modelo tradicional, no qual se deve **transmitir – verificar – registrar**, adotando uma postura de mediação, nitidamente oposta a uma escala predefinida e estática que deve ser alcançada. Por esse motivo, essa avaliação tem sido chamada de mediadora ou, em contexto um pouco distinto, de **avaliação emancipatória**. Seja qual for a terminologia adotada, trata-se de uma perspectiva de avaliação comprometida não apenas com o processo de ensino-aprendizagem, mas também com a transformação social, entendendo a educação como um instrumento não só de colocação de trabalhadores no mercado de trabalho, mas sua inclusão efetiva na complexidade social, política, econômica e cultural, comprometida com a construção de um mundo mais justo, solidário e ético.

As diferentes tradições pedagógicas, hegemônicas ou não, podem adotar diferentes arquiteturas educativas (veja **texto 12 - Resolução de Problemas e Novas Arquiteturas Educativas**), tendendo a deslocar o seu centro de gravidade em direção a diferentes elementos do processo de ensino-aprendizagem. Quando o centro de gravidade está deslocado em direção ao professor ou mesmo ao conteúdo conceitual, é comum que seja avaliado o que é fácil de ser medido, e assim, frequentemente, a aprendizagem passa a ser objeto de verificação centrada em informações factuais e não em conceitos e na sua aplicação em contexto. No entanto, os testes padronizados de larga escala, por vezes de âmbito nacional, baseados no conhecimento de fatos, não podem medir os processos e as habilidades que nossos alunos precisam desenvolver para participar do mundo atual e de sua transformação. Esses instrumentos também não permitem avaliar de que maneira os estudantes sabem aplicar o que aprenderam para melhorar as suas próprias vidas e as de suas comunidades.

Isso não significa, no entanto, que os testes padronizados sejam inúteis e não cumpram a sua finalidade, em especial no ensino médio, quando eles podem sinalizar o provável sucesso (ou não) dos alunos como calouros em instituições de ensino superior. No entanto, eles não são suficientes para sinalizar o progresso dos alunos em uma dimensão mais ampla. Para que se possa desenvolver instrumentos autênticos e precisos para avaliar atividades educativas nas quais os alunos se envolvem ativamente, realizando pesquisas, elaborando novas ideias e hipóteses, é necessário confiar a tarefa aos professores que os conduzem dentro das salas de aula. Essa afirmação provém de uma das maiores autoridades no ensino de ciências da atualidade e se contrapõe à tendência generalizada de supor que apenas avaliações externas e de larga escala possam adequadamente aquilatar a qualidade da educação:

> Apenas na educação nossa sociedade assume que o julgamento dos profissionais diretamente envolvidos em suas atividades (*front-line professionals*) deva obrigatoriamente ser validado por uma terceira parte – neste caso, por elaboradores de testes em grandes quantidades. Aprendizagem autêntica pode ser aquilatada de maneira precisa apenas pelos professores que trabalham diretamente cm os alunos a serem avaliados (Thier& Daviss, 2001, p. 115).

Quando o centro de gravidade da atividade educacional está deslocado em favor do problema a ser investigado, demandando ações de elaboração e teste de hipóteses, em especial

nos chamados **projetos de ensino**, nos quais os alunos trabalham intensamente diante de problemas concretos realizando trabalhos práticos, a avaliação e a atividade instrucional devem caminhar juntas. Trata-se, portanto, de uma avaliação inevitavelmente processual, que deve estar "embutida" no próprio processo educacional. As boas práticas avaliativas, definidas em um documento estadunidense, deveriam começar por definir de maneira operacional o que é realmente importante aprender e os métodos utilizados para que os professores possam quantificar aquilo que o aluno deve aprender, por exemplo, ao definir que os alunos devem saber planejar uma investigação. Os alunos sabem antecipadamente o que se espera deles e o que podem encontrar nos rituais de avaliação.

As atividades planejadas para servir de instrumentos de avaliação devem se parecer o mais possível com as atividades cotidianas dos estudantes, evitando-se assim a necessidade de desenvolver habilidades referentes à destreza de responder a testes padronizados, o que de fato interrompe o aprendizado verdadeiro. Essa observação, novamente retirada daquela importante referência bibliográfica (Thier & Daviss, 2001, p. 115), permite-nos perceber um erro muito comum nas experiências de inovação curricular. Como veremos adiante, testes padronizados têm sido apontados erroneamente como indicadores objetivos e isentos de sucesso no desempenho acadêmico dos estudantes, e não faz sentido adotar novas arquiteturas educativas se forem mantidas as velhas estruturas de avaliação para monitorar o seu possível sucesso. Os testes padronizados em larga escala demandam treinamento específico, destrezas mecânicas incompatíveis com propostas educacionais mais profundas e consistentes que visam a desenvolver habilidades relacionadas com a investigação científica.

Estudo realizado com estudantes de ensino médio de São Paulo revelou que o desempenho em questões de Biologia do ENEM depende em larga medida do treinamento para enfrentar um teste longo, no qual o gerenciamento do tempo é crucial para o seu sucesso. Alunos que não leram o enunciado de uma amostra de questões do Novo ENEM tiveram nível de acerto igual ou SUPERIOR ao daqueles que receberam as questões completas, comprovando que um importante elemento de sucesso nesses testes padronizados em larga escala é o treinamento, em especial saber descartar informação que não é relevante para responder à questão. Para ter acesso ao estudo completo, **clique aqui**!

13.5 Instrumentos de Avaliação em Larga Escala

A partir da década de 1990, o Brasil passou a monitorar a qualidade da educação a partir de instrumentos de avaliação em larga escala (ALE), aplicados a milhares de estudantes em curto espaço de tempo. Em outros países, como os Estados Unidos, tais avaliações já eram utilizadas há muito tempo e isso explica a contundência da crítica da referência bibliográfica estadunidense citada na seção anterior. Para compreendê-la, no entanto, é necessário entender a organização básica das ALEs.

Aplicar uma prova simultaneamente a sete milhões de estudantes, como é o caso do ENEM, exige o recurso a questões de resposta objetiva, que tenham sido selecionadas de acordo com as finalidades da avaliação. Entre 1998 e 2008, o ENEM esteve baseado na chamada Teoria Clássica de Testes (TCT), referida por alguns autores como Teoria Clássica da Medida (TCM). Trata-se de uma prova pontual e criterial, que permite classificar os estudantes de acordo com critérios previamente estabelecidos. No entanto, por suas características, a prova não permite comparações de um ano a outro, assim como no caso dos vestibulares. Se a média dos alunos aumenta de um ano para o outro, duas hipóteses são igualmente válidas: a primeira diz que os alunos se apresentaram com maior aptidão para responder às questões; a segunda diz que as questões ficaram mais fáceis. Apenas com os resultados do exame, é impossível testar qualquer uma das duas hipóteses e qualquer afirmação sobre a melhora da qualidade da educação (ou maior facilidade do exame) deverá ser vista com grande reserva.

Uma das vantagens das ALEs baseadas na TCT é a sua legitimidade. Os alunos podem conferir suas respostas e fiscalizar as consequências do ranqueamento, como a distribuição de vagas em universidades públicas, a distribuição de bolsas de estudo etc. Uma das desvantagens de tais provas é, justamente, o fato de ser impossível comparar resultados de diferentes aplicações da prova. Os vestibulares brasileiros e diversos testes aplicados em outros países, como WAIS e Raven (testes para determinação do Quociente de Inteligência), se valem de testes desse tipo.

A partir da década de 1960, uma nova possibilidade concreta se abriu com a disponibilidade de recursos computacionais, a partir da Teoria de Resposta ao Item (TRI). Segundo ela, é possível diferenciar candidatos que acertaram o mesmo número de questões se o grau de dificuldade delas for conhecido e diferente. ALEs baseadas na TRI permitem, adicionalmente,

comparação entre aplicações em momentos diferentes, uma vez que o grau de dificuldade média do exame pode ser calibrado. Essa é uma das vantagens desse tipo de teste. Entre as suas desvantagens, figura o fato de as questões passarem por uma pré-testagem, ou seja, são aplicadas em uma população o mais semelhante possível com a que deve ser avaliada. Isso implica que nenhuma questão é rigorosamente inédita, tendo já ocorrido caso de as questões pré-testadas terem sido copiadas pela escola.

A prova do ENEM 2011 trazia questões que tinham sido utilizadas em um simulado numa escola particular do Ceará duas semanas antes da prova. O caso foi muito ruidoso e conduziu a inquérito policial, que acabou por indiciar pessoas ligadas ao Colégio Christus, de Fortaleza. A apuração revelou que as questões foram pré-testadas naquela escola, e foram inadvertidamente copiadas de maneira criminosa. **Clique aqui** para ver a notícia!

Outra desvantagem dos exames baseados na TRI refere-se à sua legitimidade. O fato de o aluno anotar suas respostas não significa que possa monitorar o seu desempenho relativo, uma vez que os resultados são normalizados, segundo critérios estatísticos, até anulando acertos considerados não indicadores da aptidão do estudante, mas apenas do acaso. Os chamados **chutes** podem ser identificados quando se conhece bem o grau de dificuldade da questão e a medida de aptidão do estudante.

O Sistema de Avaliação da Educação Básica (SAEB) utiliza testes desse tipo desde a década de 1990, o que permite monitorar a variação das médias de desempenho acadêmico dos estudantes ao longo do tempo.

A crítica de Herbert Thier à **avaliação feita por máquinas**, em sua defesa da necessidade de ouvir os professores que trabalham diretamente com os estudantes em sua avaliação, refere-se à outra modalidade de ALE, chamada Avaliação Adaptativa Computadorizada ("CAT", na sigla em inglês). Trata-se de um instrumento baseado na TRI, em que cada avaliação pode ser diferente para cada aluno, conforme o seu nível de proficiência.

A **Figura 13.2** procura mostrar a ideia central da avaliação adaptativa, que supõe inicialmente que a maioria dos estudantes tem um nível médio de proficiência como reflexo de sua aptidão. Assim, supondo-se que o nível médio seja de 100 pontos, serão apresentadas questões com nível médio de dificuldade. O acerto tenderá a fazer com que sejam apresentadas questões mais difíceis, mas o erro vai fazer com que questões mais fáceis sejam apresentadas até que

haja estabilidade dos acertos em certo nível de dificuldade de questões. O erro estimado (linha vermelha) associado é, em geral, muito grande.

A partir da primeira resposta, o computador seleciona a questão seguinte baseado na suposição de que um acerto indica potencial de proficiência superior ao médio, ao passo que um erro indica potencial de proficiência inferior ao médio. A seleção depende de um algoritmo, que se deve basear em teorias estatísticas, que pressupõem que o desempenho do estudante se estabiliza ao longo do teste, o que faz diminuir o erro da medida com o aumento do número de questões corretamente respondidas em um mesmo nível de proficiência. Assim, ao contrário dos exames tradicionais, os exames CAT têm número de questões variável, conforme o desempenho do estudante e o objetivo do exame.

Os exames computadorizados têm sido alvo de discussão intensa. De maneira geral, eles permitem eliminar algumas fontes de risco como, por exemplo, de sigilo. Na prova do ENEM de 2009, o furto de um único caderno de questões implicou a anulação da prova, causando um prejuízo estimado em R$34 milhões, em valores da época. Seria pouco provável que o MEC conseguisse disponibilizar sete milhões de computadores ao mesmo tempo para a realização do ENEM; assim, com a TRI, as provas podem ser realizadas assincronamente, bastando para isso que o banco de questões seja suficientemente amplo, tanto maior quanto maior for o universo de interessados, e adequadamente pré-testado. A grande vantagem da avaliação adaptativa é permitir uma medida de aptidão com o mesmo erro para todas as faixas de proficiência, ao passo que as avaliações padronizadas têm erro menor para a faixa de desempenho médio, sendo maior para os extremos à direita e à esquerda da curva normal.

As desvantagens das avaliações adaptativas referem-se, principalmente, à calibragem das questões. Do ponto de vista ideal, os itens a calibrar devem ser administrados em meio a outros itens que não serão utilizados na montagem do exame, o que requer um razoável esforço logístico e, mesmo assim, a margem de erro da calibragem dependerá, ela própria, de parâmetros a serem considerados entre o grupo-alvo da pré-testagem e o grupo de estudantes a ser examinado. Grupo heterogêneo de estudantes a avaliar, como é comum em exames nacionais, sobretudo em um país como o Brasil, implica erro maior na calibragem das questões.

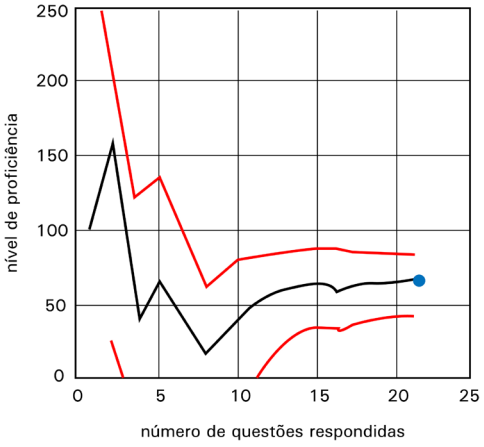

Figura 13.2: Variação e estabilização do desempenho de um estudante em uma avaliação adaptativa computadorizada (CAT). O eixo x representa a quantidade de questões resolvidas e o eixo y, o nível de proficiência.

Outra desvantagem refere-se ao fato de que os itens de dificuldade maior ou menor requerem uso monitorado, pois a probabilidade de utilização múltipla é maior, o que traz problemas para exames em larga escala, pois essas questões se tornam conhecidas, o que invalida a calibragem inicial e introduz questões éticas e de segurança.

Diante desses problemas, outra desvantagem aparece, expondo a sofisticação dos recursos exigidos, sobretudo pela CAT. É necessária uma pré-testagem rigorosa, de grande envergadura e constante, acompanhando as variações de aptidão média do grupo que procura a avaliação, um banco de questões muito grande e diversificado, e infraestrutura sofisticada, o que inclui software com algoritmo ajustado para seleção de itens, de modo a evitar superexposição de itens distante da média, e hardware adequado à demanda.

13.6 Avaliação de Desempenho e o Estatuto Escolar do Erro

O estudo da problemática da avaliação deve ter trazido à tona as possíveis relações entre o que foi discutido neste texto e no texto 9 (**Estatuto Escolar do Erro e das Concepções Alternativas dos Estudantes**), em que foi discutida a necessidade de se conferir uma nova dimensão ao erro. Naquela oportunidade, foi ressaltada a importância de se entender o erro como possível expressão de concepções que os estudantes constroem ativamente e que cabe à escola modificar. Isso valoriza uma pluralidade de concepções e abre novas perspectivas para as atividades de avaliação.

A partir do estudo das chamadas concepções prévias dos estudantes, é possível planejar estratégias de avaliação, nas quais as diferentes possibilidades de expressão dos estudantes estejam presentes, permitindo ao professor estimar em que medida as atividades escolares contribuíram para a sua possível modificação. Isso significa que é necessário conhecer as concepções dos estudantes **antes do início** da sequência didática planejada, o que implica conhecer a bibliografia da área de pesquisa de ensino da temática em estudo, valendo-se de avaliação diagnóstica, inclusive. Vimos exemplos diversos, como no caso da fotossíntese.

Ao mesmo tempo, é importante assegurar que a diversidade de concepções esteja presente nas possibilidades apresentadas pelos instrumentos de avaliação. Por exemplo, pode-se apresentar uma ilustração, que apresenta o desenho de um jato d'água emergindo de um furo em uma garrafa

cheia de água, ou uma explicação para a respiração das plantas e sua relação com a atividade fotossintética. Assim, em lugar de apresentar opções absurdas ou pouco factíveis como alternativas, até mesmo em questões de múltipla escolha, é possível fazer da avaliação uma parte **embutida** das atividades didáticas de uma ampla variedade de situações, de cunho prático, inclusive.

13.7 Recapitulando

- A concepção de avaliação está intimamente associada ao modelo pedagógico adotado, expressando um conjunto de valores em relação à educação e à sociedade;
- As avaliações podem ser de diferentes tipos, em função da perspectiva que se adote em relação ao momento de sua aplicação, escala de referência etc.;
- Modelos pedagógicos próximos de práticas tradicionais valorizam estratégias de avaliação pontual, somativa e normativa;
- Modelos pedagógicos contra-hegemônicos se valem da avaliação dita mediadora ou emancipatória, na qual se valorizam variáveis que tendem a promover o aprendizado, valorizando estratégias processuais e formativas;
- Arquiteturas pedagógicas inovadoras devem ter estratégias de avaliação igualmente inovadoras e não devem ser avaliadas por instrumentos tradicionais, tais como testes padronizados;
- As avaliações em larga escala se valem de diferentes metodologias e estão baseadas em diferentes teorias estatísticas, como a Teoria Clássica dos Testes e a Teoria de resposta ao Item.

Referências

BIZZO, N. **Metodologia de Ensino de Biologia e Estágio Supervisionado**. São Paulo: Ática Educadores, 2013.

FERNANDES, P.G.M. **Sistema computadorizado de avaliação adaptativa em larga escala (SCAALE)**. Brasília - UNB: Monografia, Departamento de Ciência da Computação, 2009.

HOFFMANN, J. **Avaliação mediadora:** uma prática em construção da pré-escola à universidade. Porto Alegre: Mediação, 2001.

LUCKESI, C.C. **Avaliação da aprendizagem escolar:** estudos e proposições. São Paulo: Cortez, 2011.

SAUL, A.M. **Avaliação emancipatória:** desafios à teoria e à prática de avaliação e reformulação de currículo. São Paulo: Cortez, 2000.

SAVIANI, D. **História das ideias pedagógicas no Brasil.** Campinas: Autores Associados, 2008.

THIER, H.D.; DAVISS, B. **Developing Inquiry-Based Science Materials:** a guide to educators. New York: Teacher College Press, 2001.