

AVALIAÇÃO COM O RECURSO A TECNOLOGIAS DE INFORMAÇÃO

CARLA LOPES¹

DANIEL MOURA²

A avaliação não deve ser vista como um mero processo de medição mas antes como parte do próprio processo de formação. Sendo a avaliação necessária, esse momento deverá ser rentabilizado e encarado como mais uma oportunidade de aprendizagem. A docência universitária colocou, durante largas décadas, a ênfase no processo de ensino centrado em aulas expositivas de um programa a ser cumprido. Nessas circunstâncias, a avaliação equivalia a um exame do que havia sido assimilado através de provas tradicionais e de notas classificatórias (Massetto, 2003). Processo distinto é o da avaliação que mede a aquisição de conhecimentos e capacidades tem como funções verificar a evolução do estudante ao longo do ano, diagnosticar dificuldades, dar retroinformação aos estudantes sobre a

¹ Professora Associada com Agregação da Faculdade de Medicina da Universidade do Porto; investigadora do Instituto de Saúde Pública da Universidade do Porto; Investigadora da Unidade de Investigação e Desenvolvimento Cardiovascular da FMUP;

² Professor Catedrático da Faculdade de Medicina da Universidade do Porto; investigador do Instituto de Farmacologia e Terapêutica da FMUP;

sua aprendizagem, avaliar a eficácia da disciplina e motivar para o estudo (Cannon, 2000).

Uma avaliação mais formativa não despende menos tempo e tem a vantagem de fornecer informações, identificar e explicar erros, sugerir interpretações quanto às estratégias e atitudes dos estudantes, estimulando assim a ação pedagógica (Perrenoud, 1999). Alguns dos métodos sugeridos para se avaliar de modo contínuo e processual passam pela observação dos estudantes durante o processo de ensino, desenvolvimento de projetos, leitura crítica e discussão de artigos, resolução de problemas, execução de testes objetivos, avaliação de casos, redação de relatórios tutelados e construção de portefólios que reúnam sistematicamente produções dos estudantes, ou ainda escalas de autoavaliação (McAleer, 2001; Tormey, 2014).

É utópico pensar-se que existe consenso sobre o melhor método de avaliar e nem sempre é fácil a aplicação de determinados métodos de avaliação, em particular quando trabalhamos com grandes números de estudantes. No entanto, é imperativo que o processo de avaliação evolua e acompanhe as alterações do processo de ensino–aprendizagem orientado para a aquisição de competências (McAleer, 2001). O recurso a novas tecnologias de informação no processo de ensino-aprendizagem tem crescido nas últimas décadas, pelo que o uso de computadores se tornou generalizado no processo de formação em todas as áreas do conhecimento. Contudo, o seu

uso para avaliação da aprendizagem está ainda longe de poder ser considerado otimizado.

Na maioria das instituições de ensino superior os testes em papel continuam a ser os mais comuns ou mesmo as únicas formas de avaliação da aprendizagem dos estudantes.

1. AVALIAÇÃO ELETRÓNICA (E-ASSESSMENT)

O termo e-assessment tem vindo a impor-se na gíria corrente pela semelhança com o e-learning, podendo definir-se como o uso de tecnologias de informação em qualquer atividade de avaliação. Algumas definições procuram ser mais precisas, como, por exemplo, a do Comité Conjunto dos Sistemas de Informação (The Joint Information Systems Committee, JISC) e da Autoridade das Qualificações e dos Currículos (Qualification and Curriculum Authority) do Reino Unido, para os quais "... e-assessment is the end-to-end electronic assessment processes where ICT is used for the presentation of assessment activity, and the recording of responses. This includes the end-to-end assessment process from the perspective of learners, tutors, learning establishments, awarding bodies and regulators, and the general public" (The Joint Information Systems Committee, 2007). O desenvolvimento das tecnologias de informação colocou ao nosso dispor diversos produtos tecnológicos sofisticados, computadores ou outros aparelhos de telecomunicações, que possibilitam a avaliação do que os estudantes estão a aprender em níveis de detalhe muito específico. Essa

avaliação poderá se feita à distância, através de simulações de situações reais e de formas que se distinguem das atividades de aprendizagem.

Se partilharmos a ideia de que a tecnologia por si só não significa uma melhoria na avaliação da aprendizagem, concordamos que o uso de novas tecnologias no processo de avaliação abre novas perspectivas e possibilidades ao processo de ensino-aprendizagem mais centrado no estudante e orientado para a aquisição de capacidades.

O processo de avaliação deve ser baseado em conhecimentos modernos sobre os processos cognitivos e a sua medição, deve ser integrado com os programas curriculares e deve ser informativo e incentivador do alcance dos objetivos por parte dos estudantes (National Research Council, 2001). Melhorar o processo de avaliação passa por uma programação cuidada que faça a conexão com 3 elementos fundamentais no processo: a teoria cognitiva, a observação e o processo de interpretação. Os estudos da cognição modernos procuram construir modelos gerais da forma como as pessoas representam os conhecimentos e se tornam aptas numa determinada matéria. A observação é o conjunto dos desempenhos possíveis que uma pessoa pode ter quando é submetida a uma avaliação: por exemplo as alternativas corretas ou os distractores que ela escolhe num teste de escolha múltipla. Uma avaliação é um pequeno mundo estruturado de forma a obter um conjunto determinado de observações. Importa que essas observações sejam de alguma forma expressões do conhecimento que

a pessoa assimilou e da aptidão que construiu. A tentativa fundamentada de explicar os resultados do desempenho de um sujeito numa prova de avaliação através da representação cognitiva que ele tem e da aptidão que ele criou tem a designação de processo de interpretação no documento do National Research Council de 2001 com o título muito sugestivo de "Knowing what students know" (National Research Council, 2001). Este processo de interpretação é em grande parte resultado do tratamento estatístico e deve muito do seu avanço aos progressos dos estudos de medida aplicados à psicologia e à educação. Uma avaliação é pois uma oportunidade para se fazer a prospeção não só do que se respondeu mas do que se está a aprender, e como. A utilização das tecnologias de informação no processo de avaliação oferece um potencial enorme para desenvolver e aplicar no quotidiano essa tentativa de interpretar os resultados de uma avaliação (a observação) em função da cognição do examinado, quer seja em autoavaliação, avaliação formativa ou sumativa.

2. VANTAGENS E DESVANTAGENS DO RECURSO A TECNOLOGIAS DE INFORMAÇÃO PARA FAZER AVALIAÇÕES

A partir do exemplo concreto de avaliações usadas em Medicina é possível fazer uma enumeração sugestiva de vantagens e desvantagens do recurso a meios informáticos, que se pode extrapolar para a generalidade das áreas do conhecimento (Cantillon, Irish et al., 2004; Hols-Elders, Bloemendaal et al., 2008).

Vantagens:

- Os estudantes têm a possibilidade de escolher o momento em que querem ser avaliados
- A informação do resultado é imediata
- A análise estatística das questões do teste pode ser feita de forma rápida e exaustiva
- A possibilidade de reutilização das questões faz-se em boas condições
- Os professores podem seguir com facilidade a evolução do desempenho dos estudantes individualmente
- Podem usar-se ferramentas (por exemplo, vídeos, gráficos, imagens 3D, simulações,...) que não são possíveis nos testes em papel
- A popularidade entre os estudantes do uso de instrumentos digitais é grande
- O aumento da qualidade de futuros testes é mais fácil
- O processo de assinalar as questões é automático e rápido e não enviesado por erros humanos, o que o torna mais válido
- Os testes podem ser escolhidos rapidamente a partir um banco de questões previamente preparado.

Desvantagens:

- Os testes estão sobretudo adaptados às questões fechadas, embora se tenha vindo a assistir à evolução dos métodos para avaliação de questões abertas
- Os estudantes com limitações técnicas no uso de computadores têm uma dificuldade artificial na realização das provas

- Os problemas de segurança para as avaliações sumativas são complexos
- O dispêndio de tempo na criação das avaliações é maior
- As condições técnicas exigidas são mais difíceis e dispendiosas
- Há necessidade de recursos humanos especializados e treinados no desenho dos testes e no armazenamento dos dados para construção de bancos de itens.

3. DESENHO DOS ITENS

Sejam gerais ou individualmente ajustados, os testes dependem sempre do desenho e da escolha cuidadosa do tipo de questões a incluir. Uma revisão de 2006 (Saclise & Gifford 2006) definiu 28 diferentes tipos de formatos de itens baseados em 7 categorias: **1)** Escolha múltipla (verdadeiro/falso, escolha alternativa, escolha múltipla convencional e escolha múltipla com novos distractores multimédia); **2)** Seleção/Identificação (verdadeiro/falso múltiplo, sim/não com explanação, respostas múltiplas, e escolha múltipla complexa); **3)** Reordenar/Rearranjar (emparelhar, categorizar, ordenar e sequenciar, organizar uma sequência de prova); **4)** Substituição/Correção (interlinear, identificação de incorreções, desenho de figuras/gráficos, correção de erros em imagens); **5)** Completar (construção numérica simples, resposta curta ou completar frase, procedimento fechado, completar matriz); **6)** Construir (escolha múltipla aberta, construção de figuras, gráficos ou mapas conceptuais, composição); **7)** Apresentação/Portfólio (projeto, demonstração/experiência, discussão/entrevista, diagnóstico/ensino).

4. TESTES ADAPTATIVOS E NÃO ADAPTATIVOS

A velocidade de cálculo dos computadores tornou possível que, nas condições normais do trabalho, se usem métodos quantitativos valiosos para a deteção de fragilidades na forma como as avaliações são feitas e, em parte, para a sua correção. Há três defeitos das avaliações que podem ser atenuados através de métodos estatísticos com cálculos feitos por computador: a inconsistência de algumas perguntas, o viés dos juízos que o avaliador faz sobre a qualidade das perguntas que ele próprio criou e a ausência de ajuste das provas à qualidade dos examinandos, quando são todos avaliados da mesma maneira. A entrevista seria, em abstrato, uma solução óbvia, mas a sua aplicação prática agrava alguns desses erros e tem problemas próprios. O seu valor está na possibilidade de ajustar e corrigir o conteúdo das avaliações à medida que ela se vai desenrolando: a inconsistência de uma pergunta pode ser detetada e esclarecida no decurso do diálogo, a exigência da avaliação pode ajustar-se ao decorrer da prova, o viés pode ser rebatido pelos argumentos do examinando. É, porém, claro que o tempo é escasso e que o viés do avaliador aumenta muito, podendo com facilidade colocar em risco a equidade das avaliações. É possível porém transferir uma boa parte do trabalho do entrevistador para o computador.

Para que o computador possa atuar como um entrevistador eletrónico que se vá adaptando às respostas que vão sendo dadas pelo entrevistado é necessário que esteja programado com uma grande base de perguntas e com um algoritmo para as escolher (Tao Y-H, et al., 2008). A qualidade das

perguntas é fundamental. O seu conteúdo tem de ser correto e relevante para o assunto da prova e o seu enunciado deve ser cuidadoso. O problema da validade do conteúdo é muito específico e não será aqui tratado. A construção de enunciados é também alvo de regras práticas bem estabelecidas e disponíveis em vários manuais ou guiões de boas práticas para a construção de itens. O aspeto da boa qualidade que se prende mais diretamente com o uso de instrumentos informáticos é a sua caracterização estatística.

Usam-se dois modelos estatísticos gerais para análise psicométrica ou edumétrica das provas e dos seus componentes, designados por teoria clássica dos testes e por teoria da resposta a itens (Hambleton, 1993; De Champlain, 2010). Com eles é possível introduzir processos independentes do juízo subjetivo do avaliador para verificar a consistência das perguntas e para fazer uma análise da qualidade estatística das perguntas.

A teoria clássica dos testes e a teoria da resposta a itens têm fundamentos matemáticos diferentes. Só a segunda permite a realização dos testes adaptados no momento ao desempenho do avaliado, mas ambas preenchem uma necessidade anterior muito importante: obter indicações sobre a dificuldade, a consistência e a discriminação das perguntas.

A teoria clássica dos testes usa principalmente regressões e correlações lineares para analisar a consistência das perguntas. Os valores da correlação

bisserial ou ponto bisserial ou o agregado de correlações do teste do alfa de Cronbach avaliam se os acertos numa dada pergunta são consistentes com os acertos no teste global ou com os acertos nas outras perguntas: o grupo de indivíduos com bom desempenho global num teste deve ter uma boa percentagem de acerto na pergunta x enquanto que o grupo de indivíduos com mau desempenho global deve ter uma baixa de acerto nessa pergunta. Deve portanto haver uma reta que se ajuste aos valores crescentes de uma e outra medição: quanto melhor é o desempenho dos indivíduos numa das perguntas tanto maior deve ser o desempenho nas outras e no global.³ Consistência e discriminação ficam assim definidas na teoria clássica dos testes fundamentalmente por ajustes a equações lineares. A análise da dificuldade exprime-se muito simplesmente pela percentagem de acertos. A exploração cuidadosa destes indicadores é um instrumento precioso para avaliar a qualidade ou falta dela das questões de um determinado teste, mas não permite construir testes individualizados para o futuro. Outra forma de análise estatística a que se chamou teoria da resposta ao item permite também fazer a análise da dificuldade, da consistência e da discriminação das questões e usar esses indicadores para a realização de testes individualizados ou adaptativos. O princípio geral da teoria da resposta ao

³ A teoria clássica dos testes caracteriza-se assim pela análise de valores de r , que não podem ser negativos, têm de ser superiores a 0 e se devem aproximar de 1, e pela análise do significado estatístico do valor de r expresso pelo valor de P inferior a uma probabilidade de erro aceitável, por exemplo 5%. A análise da correlação linear permite também definir a inclinação da reta que deve ser a maior possível para que a pergunta se considere discriminativa, isto é, a pequenas diferenças entre o desempenho dos indivíduos correspondem grandes diferenças nos acertos à pergunta.

item é o da regressão não linear ou seja, o ajuste dos dados é feito a curvas definidas por uma equação não linear que se julgue adequada para descrever o que se está a passar (Downing, 2003; Ayala, 2009). Há aqui uma associação de dois conceitos com o mesmo nome, um matemático e outro psicológico (Embretson & Reise, 2000).⁴ A forma habitualmente mais simples de expor a base da teoria da resposta ao item é explicar a curva característica do item.⁵ De forma pragmática analisa-se se o item é consistente se obedecer à condição de que quanto maior for a aptidão de um indivíduo maior é a probabilidade de acertar na resposta correta ao item. Se se dispuser de um número grande de indivíduos e um número grande de itens numa prova pode tomar-se a avaliação da aptidão pelo desempenho geral da prova. Na linguagem quotidiana diz-se que quem tem boa nota deve acertar muito à pergunta x. Se não for assim, a pergunta não é coerente e a sua qualidade fica posta em causa.⁶ Feita a análise termina-se, assim, com dois marcadores para cada item: dificuldade e discriminação. Estes valores permitem já colocar o item numa grande base de dados preparada para testes adaptativos. A ideia geral para estes testes é a de que

⁴ É o traço latente, isto é, o que não se vê mas está a influenciar o que se passa. Para a psicologia o traço latente que influencia o acerto numa determinada prova ou questão é a aptidão ou a capacidade crucial para resolver aquele tipo de problema. Para a matemática é o valor dos parâmetros da equação que melhor explica os dados.

⁵ É a curva, com a forma de sigma, que relaciona o traço latente do indivíduo expressa em desvios-padrão da média no eixo horizontal e a probabilidade de acerto ao item no eixo vertical (Embretson & Reise).

⁶ Os valores matemáticos para definir a coerência do item são o ajuste de curva e o seu significado estatístico. Como se avalia a dificuldade e a discriminação? A primeira pela posição da curva mais à esquerda ou mais à direita ou pelo valor da aptidão necessário para que a probabilidade de acerto seja de 50%. A segunda pela inclinação do sigmoide: quanto mais verticalizado for maior é a discriminação (idem).

se deve examinar os indivíduos muito aptos com itens difíceis e examinar os poucos aptos com itens fáceis, isto é, novamente na linguagem quotidiana, a prova deve ser feita ao nível do estudante (Weiss, 1982; Weiss, 2004). Parece que estaríamos num ciclo vicioso sem saída porque o que se procura nas avaliações é o contrário: não se conhece o nível do estudante e é para isso que ele é examinado. O algoritmo matemático e o poder de computação resolvem o problema: as perguntas iniciais procuram “tomar o pulso” à preparação do estudante e a partir do desempenho inicial faz-se uma inferência aproximada. O programa passa então a usar uma primeira seleção larga de perguntas ao nível da aptidão estimada. Cada resposta determina a oscilação da dificuldade da pergunta a escolher a seguir. Em determinado ponto do decurso do teste está encontrado o melhor valor possível da aptidão do estudante.

A dificuldade maior está na construção do banco de questões: têm de ser muitas (aproximadamente um milhar), boas, calibradas e cobrir todo o espectro de dificuldade.⁷ O assunto de que trata o item tem de ser relativamente estável para que as perguntas possam ficar calibradas durante bastante tempo. Os grandes sistemas de testes para aptidões gerais desenvolvidos para os estudantes do ensino secundário nos EUA, ou para as provas de ingresso nos cursos de MBA, ou para o licenciamento da prática de farmacêutico, ou ainda, surpreendentemente, de certificação parcial do domínio de uma língua estrangeira, são exemplos de testes adaptativos que se podem comprar já feitos e prontos a usar. Em domínios especializados, se

⁷ De -3 a +3 desvios-padrão da média.

o professor tiver no mínimo cerca de 100 estudantes e pacientemente criar, calibrar e colecionar perguntas boas, pode construir um banco de itens que a partir de certa dimensão pode estar em condições de se usar em testes adaptativos. Torna-se necessário que a instituição tenha uma organização física que permita salas e computadores e disponibilize o programa que faça a gestão do banco de perguntas em tempo quase imediato. Não é fácil e é dispendioso mas tem vindo a ser cada vez mais acessível e é já praticável em condições reais (Amaral, Moura et al., 2008).

Adivinha-se também o próximo passo que encerra o círculo: não é irrealista pensar-se que a análise de itens permitirá identificar os fatores cognitivos importantes para um determinado tipo de bom desempenho e isso pode identificar de forma mais clara o que necessita de melhor e mais esforço de estudo e de ensino (Hambleton, 2000).

5. IMPLEMENTAÇÃO DO PROCESSO

A implementação deste processo de avaliação eletrónica exige requisitos específicos relacionados com a coordenação, administração, suporte técnico, segurança e preparação dos avaliadores e dos avaliados (Authority, 2007; Hols-Elders, Bloemendaal et al., 2008), que vai para além do âmbito deste capítulo. Referimos apenas de forma resumida algumas das sugestões que devem ser consideradas e que passam por: **a)** tornar explícito e comunicar quais os benefícios para a instituição, processo que tem que ser conduzido pelos próprios gestores da instituição; **b)** assegurar as infraestruturas relativas a equipamento e programas informáticos e espaços físicos

apropriados; **c)** assegurar a construção de um banco de questões com qualidade; **d)** reconhecer nas diferentes áreas as potencialidades do e-assessment que os exames em papel não têm; **e)** assegurar a segurança nos testes sumativos; **f)** estar preparado para os possíveis problemas tecnológicos que podem surgir no decorrer do processo; **g)** assegurar a informação e a formação de professores e estudantes.

Deve ainda referir-se que o problema da segurança é particularmente complexo, não só para evitar a fraude, mas também para proteger os itens da familiarização indevida. A divulgação de um item altera por completo a sua dificuldade e a sua discriminação e torna inútil a calibração que ele tinha quando foi incluído no banco de dados.

A implementação do processo em qualquer instituição requer uma avaliação prévia de eficiência e custos, dependendo de especificidades como a área de ensino, o número de alunos, recursos humanos e materiais, etc. Para além disso é necessário o envolvimento prévio de uma equipa especializada capaz de assegurar a manutenção técnica e logística do processo.

A falta de conhecimento e domínio das estratégias de conceção e de aplicação dos testes adaptativos, a falta de tempo (ou mesmo o desinteresse) para adquirirem esses conhecimentos e a falta de recursos materiais são algumas das razões apontadas para a não implementação generalizada do processo. É claro que a sugestão de substituir os tradicionais testes de papel e caneta pela avaliação por recurso ao

computador não é tarefa fácil, em particular com grande número de estudantes. Contudo, parece um processo inevitável, que terá certamente um incremento nos próximos anos, e que ajudará à evolução do próprio processo de ensino-aprendizagem.

O conteúdo, o formato e a frequência da avaliação devem ser o resultado dos objetivos específicos dos diferentes programas curriculares. Mas as decisões sobre o uso de avaliações formativas ou sumativas, a frequência da realização e os padrões dos diferentes domínios de competência a ter em conta permanecem um desafio. Permanece por definir também a melhor forma de avaliar determinadas qualidades como profissionalismo, espírito de equipa ou outras capacidades muitas vezes difíceis de definir e quantificar.

Referências

- AMARAL, M., D. Moura, et al. (2008), *Development of resources for computer-based testing in campus-wide IT systems*. Proceedings of the IASK International Conference Teaching and Learning 2008. Teaching and Learning 2008. ISBN: 978-972-99397-8-5. pp. 300-307.
- AYALA, R. J. (2009), *The theory and practice of item response theory*, New York, The Guilford Press.
- CANNON R, N. D. (2000), *Assessing the students. A handbook for teachers in universities and colleges. A guide to improving teaching methods*, London, Kogan Page, pp. 207-231.
- CANTILLON, P., B. Irish, et al. (2004), "Using computers for assessment in medicine", *BMJ* 329 (7466), pp. 606-9.
- DE CHAMPLAIN, A., F. (2010), "A primer on classical test theory and item response theory for assessments in medical education", *Med Educ* 44(1), pp. 109-17.
- DOWNING, S. M. (2003), "Item response theory: applications of modern test theory in medical education", *Medical Education* 37, pp. 739-45.
- EMBRETSON, S. E. & S. P. Reise (2000), *Item response theory for psychologists*, Mahwah, New Jersey, LEA Publishers.
- HAMBLETON, R. K. (1993), *Principles and selected applications of item response theory*. Educational Measurement, R. L. Linn, Phoenix, AZ, Oryx Press, pp. 147-200.
- HAMBLETON, R. K. (2000), "Emergence of item response modelling in instrumental development and data analysis", *Medical Care* 38 (suppl. 6), pp. 1160-5.
- HOLS-ELDERS, W., P. Bloemendaal, et al. (2008), "Twelve tips for computer-based assessment in medical education", *Med Teach* 30(7), pp. 673-8.
- MASSETTO, M. T. (2003), *Competência pedagógica do professor universitário*, São Paulo, Summus editorial.

MCALEER, S. (2001), *Choosing assesement instruments. A practical guide for medical teachers*. J. A. Dent & R. M. Harden, London, Churchil Livingstone, pp. 303-313

MCALEER, S. (2001), *Formative and summative assessment. A practical guide for medical teachers*, J. A. Dent & R. M. Harden, London, Churchil Livingstone, pp. 293-302.

NATIONAL RESEARCH COUNCIL (2001), *Knowing what students know: The science and design of educational assessment*, Washington, DC, National Academy Press.

PERRENOUD, P. (1999), *Avaliação: da excelência à Regulação ds Aprendizagens - Entre duas lógicas*, Porto Alegre, Artmed.

SACLISE, K. & B. Gifford (2006), "Computer-based assessment in E.learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms." *The Journal of Technology, learning and assessment* 4.

SCOTTISH QUALIFICATIONS AUTHORITY (2007), "e-Assessment: Guide to effective practice", acedido a 28 de setembro de 2010, [www.sqa.org.uk].

TAO Y-H, Wu Y-L, Chang H-Y (2008), "A practical computer adaptive testing model for small-scale scenarios". *Educational Technology & Society*, 11: 259–274.

THE JOINT INFORMATION SYSTEMS COMMITTEE (2007), "Effective Practice with e-Assessment", www.jisc.ac.uk

TORMEY, W. (2014) *Education, learning and assessment: current trends and best practice for medical educators*. *Ir J Med Sci* DOI 10.1007/s11845-014-1069-4.

WEISS, D. J. (1982), "Improving measurement quality and efficacy with adaptive testing", *Applied Psychological Measurement*, 6, pp. 473-92.

WEISS, D., J. (2004). "Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education", In: *Measurement and Evaluation in Counseling and Development*, Volume 37, Number 2, pp. 70-84.