Direcção de Educação Médica da Faculdade de Medicina • Universidade de Coimbra

10
Fevereiro 2009

Essências EDUcare

Testes escritos: que formato escolher?

É um facto que a esmagadora maioria das unidades curriculares dos cursos de Medicina/Medicina Dentária incluem alguma forma de avaliação sumativa escrita dos seus alunos. O teste escrito é o método por excelência para a avaliação dos objectivos do domínio do conhecimento e, quando bem concebido, tem revelado a sua validade e fidedignidade ao longo do tempo e em contextos escolares muito diversificados. Mas será que todos os testes escritos revelam as mesmas qualidades? E que consequências tem a opção por diferentes formatos de teste?

1. De que critérios disponho para apreciar as vantagens e desvantagens de cada formato de teste?

A finalidade de qualquer avaliação é a de permitir fazer inferências acerca da proficiência dos alunos numa determinada área do saber. Todavia, torna-se evidente que é impossível avaliar tudo aquilo que é importante que o aluno domine no tempo limitado disponível, o que transforma o desafio de avaliar num problema de amostragem do conhecimento. Essa amostragem deve poder fornecer uma base para estimar a competência do aluno numa área do conhecimento mais abrangente. A natureza e qualidade da amostra que se seleccione determina a medida em que a estimativa da competência é precisa e reprodutível (fidedigna, consistente) e adequada (válida).

Se a amostra não for representativa do domínio mais amplo de conhecimento que se pretende avaliar, os resultados do teste estarão enviesados. Se a amostra for muito pequena, os resultados poderão não ser suficientemente estáveis para garantir que reflectem a competência "real" dos alunos. As principais considerações sobre os problemas de validade e fidedignidade dos testes encontram-se sintetizados nas tabelas seguinte.

Validade

- A validade de um teste é a medida em que este consegue medir aquilo que é suposto medir (eg. uma régua é um instrumento válido para medir distâncias, mas não o peso ou volume de objectos).
- A maioria das competências não pode ser observada directamente, pelo que é importante que o teste permita coligir evidências que assegurem que as inferências que fazemos são válidas (eg. num teste válido, os peritos devem, por norma, obter melhores resultados do que os iniciados; deve também ser possível proporcionar evidência de que o teste reflecte uma distribuição adequada dos conteúdos e objectivos de aprendizagem, designadamente, através de um blueprint/tabela de especificações; idealmente, deve ser possível obter informação de natureza técnica e estatística sobre a qualidade dos itens do teste considerados individualmente).
- Uma boa validação dos testes deve utilizar várias fontes de evidência.

Fidedignidade/Consistência

- Reflecte a confiança que temos de que o teste produziria a mesma seriação dos alunos se este fosse repetido, ou aplicado um teste equivalente (paralelo).
- Um teste representa, na melhor das hipóteses, uma amostra de todas as perguntas possíveis.

Para além dos critérios da validade e fidedignidade, alguns autores apontam ainda factores como o impacto educacional do teste, isto é, o modo como influenciam a abordagem mais superficial ou profunda com que o aluno adopta perante o currículo, ou a relação custo-benefício, determinada não apenas pelos custos da correcção dos testes -esforço e tempo dispendidos - mas também pelo processo da sua construção. Fica então claro que os testes que optarmos por utilizar para avaliar as aprendizagens deverão ser analisados, em contínuo, à luz dos 4 grandes critérios que aqui referimos, sendo que a produção de um dossier documental de evidências se revela um passo muito útil neste processo.

2. Que tipologias e formatos de teste escrito estão à nossa disposição?

Considerando a grande diversidade de formatos de perguntas de teste, é útil recorrer a uma categorização em função de duas grandes características principais: o formato da resposta e o formato do estímulo (do enunciado da pergunta).

Podemos então distinguir, segundo o formato da resposta, os seguintes agrupamentos de perguntas:

- Perguntas de Resposta Construída: Todas aquelas em que ao aluno é pedido que gere a sua resposta, ao invés de a seleccionar de um conjunto pré-determinado de opções (Resposta curta, perguntas de desenvolvimento (ensaio)).
- Perguntas de Resposta Seleccionada: (verdadeiro-falso simples, escolha múltipla do tipo "one best answer", verdadeiro-falso múltiplas ou complexas, extended-matching ou Key-features).

A categorização, segundo o formato do estímulo, distingue dois tipos de perguntas:

- Perguntas desprovidas de contexto: são aquelas que não revelam preocupações quanto ao cenário a criar, nem fornecem elementos de contextualização ao examinando. Apelam quase sempre à memorização estrita e são muito orientadas para a avaliação de conhecimento factual, de nível cognitivo mais básico. O conhecimento, que estas perguntas pretendem avaliar, é um requisito necessário, mas não suficiente, para avaliar a resolução de problemas e para inferir a competência médica em contexto "real".
-Perguntas ricas em contexto: Este tipo de perguntas consiste na (re)criação de um cenário ou caso e em perguntas que envolvam a tomada de decisões relativas a esse mesmo cenário. Quando, como neste caso, as perguntas são directamente relacionadas com um caso/cenário que exige a tomada de decisões, os processos cognitivos invocados são completamente diferentes daqueles evocados pelas perguntas desprovidas de contexto, sendo que a principal diferença reside no facto de, ao raciocínio linear e simples, do tipo sim-não, das últimas, corresponder um raciocínio do tipo proposicional, em que os alunos pesam um conjunto de unidades de informação ao tomarem uma decisão.

3. Formatos de teste, segundo o formato da resposta

3.1. Testes de Resposta Construída

3.1.1.Perguntas de Desenvolvimento (de Ensaio): Este tipo de perguntas são ideais para avaliar o modo como os alunos conseguem sintetizar, formular hipóteses e gerar soluções, encontrar relações, comparar e discutir semelhanças ou diferenças e aplicar procedimentos conhecidos a situações novas e não familiares. Podem também fornecer indicações quanto à competência escrita do aluno e da sua capacidade de processar a informação. Todavia, apresentam como grande desvantagem a sua habitualmente baixa fidedignidade/consistência, por exigirem maior tempo na resposta (menor ratio item/hora de teste) e pelo facto do processo da sua correcção poder ser afectada pela inconsistência intra e inter-avaliadores. Idealmente, devem definir-se critérios relativamente rígidos de correcção

(rubricas), procurando evitar os problemas relacionados com o fenómeno da inconsistêcia já referindo; a demanda por uma maior objectividade' mediante a inclusão de critérios excessivamente rígidos, levantará também alguns problemas, porquanto proporciona poucos ganhos na consistência e uma considerável perda de validade.

Este tipo de perguntas, dados os problemas enunciados e os custos de tempo e recursos envolvidos na sua correcção, deve apenas ser utilizado quando seja necessário avaliar competências que seria impossível avaliar de outra forma, tais como, a criatividade ou a capacidade de escrita, em que a geração espontânea da resposta pelo aluno é considerada essencial.

Refira-se ainda a existência de variantes deste tipo de pergunta, como a pergunta de ensaio modificada, em que a um mesmo caso ou cenário correspondem várias perguntas que dele decorrem, cronologicamente ou não. Este formato encerra alguns problemas relacionados com a dificuldade em assegurar a independência das questões.

Exemplo de pergunta de ensaio, destinada a avaliar a aplicação de conhecimento a situações novas: Durante o curso aprendeste os fundamentos do mecanismo de biofeedback do ACTH. Aplica este mecanismo ao controlo da diurese.

3.1.2. Perguntas de Resposta Curta: Este tipo de perguntas implica que o aluno gere uma resposta curta, habitualmente com não mais de uma ou duas palavras. A Resposta Curta procura conciliar características das perguntas de escolha múltipla com as perguntas de desenvolvimento, mas ao fazê-lo, existe o sério risco de não se conseguir alcançar os intentos desejados. É importante descrever quão detalhada deve ser a resposta, bem como providenciar uma chave de correcção bastante detalhada para quem tem a responsabilidade pela correcção. As perguntas de Resposta curta não são viáveis para a avaliação dos níveis básicos de conhecimento factual, por serem substituídas com vantagem pelas perguntas de escolha múltipla, atendendo a que estas últimas são mais consistentes e permitem uma amostragem mais ampla do conhecimento a avaliar. As perguntas de resposta curta apresentam as mesmas fragilidades que os ensaios, já que são pouco económicas, quer do ponto de vista da sua construção, quer da correcção.

3.2.1. Perguntas de Verdadeiro-Falso: estas consistem em afirmações relativamente às quais o aluno tem que indicar se são verdadeiras ou falsas. Este tipo de pergunta é usada frequentemente para testar o conhecimento de factos isolados, de nível básico, e permite a cobertura de uma ampla amostra de conhecimento num curto período de tempo, o que é, porventura, a única característica positiva que se lhes pode apontar. Em tudo o resto, acarretam mais problemas do que a maioria dos outros formatos de perguntas de teste: são bastante difíceis de construir sem falhas e ambiguidade, o que cria uma maior propensão para testar apenas a recordação de factos isolados; os revisores descartam muito mais frequentemente perguntas de verdadeiro/falso por imprecisões e ambiguidades ; é impossível de perceber se, quando o aluno assinala a resposta como falsa, ele saberia efectivamente qual a opção correcta alternativa e, por último, estas perguntas são muito mais susceptíveis à resposta dada ao acaso (50% de hipótese acerto ao acaso, em cada pergunta).

3.2.2. Verdadeiro-Falso múltiplas ou complexas: implicam que o aluno assinale mais do que uma opção de resposta. Pela positiva, refira-se que, embora tomem mais tempo a responder que outros formatos de resposta seleccionada, a sua consistência não é, habitualmente, muito mais baixa. Quanto às fragilidades, elas são várias e incluem a dificuldade na construção de opções que sejam absolutamente correctas ou incorrectas, bem como o processo de cotação, que se apresenta deveras complexo e intrincado.

Exemplo: Quais dos seguintes fármacos pertencem ao grupo inibidor ACE? a)atenolol; b)pindolol; c)amilorida; d)furosemida; e)enalapril; f)clopamida; g)epoprostenol; h)metoprolol; i)verapamil; j)digoxina; k)captrotil; l) propanolol

3.2.3. Escolha Múltipla clássica, do tipo "one best answer": Este é um formato de pergunta de teste clássica e muito disseminada. Consiste num tronco ou enunciado, numa pergunta (o lead-in) e num número variável de opções (mais frequentemente, uma opção correcta em cinco possíveis), sendo que o aluno assinala apenas a opção que considera a (mais) correcta. É um tipo de questão flexível e razoavelmente simples de construir (mais do que as do tipo verdadeiro-falso) e administrar, sendo muitíssimo eficiente a correcção por leitura automática/óptica. São especialmente adequadas quando se pretende uma amostragem ampla do domínio de conhecimento a avaliar e o número de examinandos é elevado. Apresentam valores de consistência elevados e, se bem construídas, podem testar mais do que a simples recordação de factos isolados, incluindo os níveis de interpretação e resolução de problemas. São apenas apontadas duas razões para a sua não utilização: quando a geração espontânea da resposta é essencial e quando o número de opções realísticas é demasiado grande. Em todos os restantes casos, as perguntas de escolha múltipla clássica são uma boa alternativa às perguntas de Resposta Construída.

Exemplo: Um rapaz com 2 anos apresenta um edema desde há 1 semana. A sua tensão arterial é de 100/60mmHg e revela edema generalizado e ascite. As concentrações séricas são: creatinina 0,4 mg/dL, albumina 1,4g/Dl e colesterol 569 mg/dL. A análise da urina revela proteínas 4+ e nenhum sangue.
Qual dos seguintes é o diagnóstico mais provável?
A. Glomerulonefrite pós-streptoccus aguda
B. Síndroma hemolítico-urémico
C. Síndroma nefrótico com lesões mínimas
D. Síndroma nefrótico por glomeruloesclerose segmentar

3.2.4. Extended-Matching: consistem numa lista de opções, numa pergunta (lead-in), e algumas descrições de casos clínicos ou outros, normalmente organizados em torno de um determinado tema ou tópico. As opções podem nunca ser escolhidas ou então ser seleccionadas mais do que uma vez para casos clínicos diferentes. Desta forma, a grande virtualidade deste tipo de perguntas emerge: a muito baixa probabilidade de resposta ao acaso, em virtude do grande número de possíveis combinações entre os diferentes casos e as opções de resposta.

Por regra, este tipo de questões exige que o aluno tome decisões baseadas num cenário ou caso relativamente complexo, pelo que são apropriadas para o nível de complexidade cognitiva mais elevado, como a resolução de problemas. Considerando que são de resposta relativamente rápida, os seus valores de consistência são normalmente elevados. São especialmente apropriadas quando se pretende fazer um número elevado de perguntas acerca de aspectos relacionados (eg. formulação de diagnósticos ou decisões terapêuticas).

Exemplo:
TEMA: Fatigue
"LEAD-IN": For each patient with fatigue, select the most likely diagnosis
OPÇÕES: A. Acute leukaemia; B. Anemia of chronic disease; C.Congestive heart failure; D. Depression ; E. Epstein-Barr virus infection; F. Folate deficiency; G.Hereditary spherocytosis; H.Hypothiroidism; J. Iron Deficiency
TRONCO(S):
1. A 19-year-old woman has had fatigue, fever and sore throat for the past week. She has a temperature of 38.3C, cervical lymphadenopathy, and splenomegaly. Initial laboratory studies show a leukocyte count of 5000/mm3 (80% lymphocytes exhibiting atypical features). Serum aspartate aminotransferase (AST,GOT) activity is 200 U/L. Serum bilirubin concentration and serum alkaline phospatase activity are within normal limits.
2 A 15-year-old girl has a two-week history of fatigue and back pain. She has widespread bruising, pallor, and tenderness over the vertebrae and both femurs. Complete blood count shows hemoglobin concentration of 7.0 g/dl, leukocyte count of 2000/mm3 , and platelet count of 15,000/mm3

3.2.5. Key-Features: é um formato ainda muito pouco explorado e que exige um grande esforço de construção. Neste tipo de questões, a uma descrição realística de um caso clínico sucede-se um pequeno número de perguntas que requerem apenas decisões essenciais ou críticas (tipicamente, pode ser colocada a seguinte pergunta: Qual dos seguintes é o procedimento mais apropriado a adoptar no imediato?). Vêm revelando validade na avaliação da capacidade de resolução de problemas e apresentam uma boa consistência. Podem ser consideradas uma variante das perguntas de escolha múltipla clássica.

Para saber mais:

- Case, S. M., Swanson, D. B. (2001) (3rd Edition). Constructing Written Test Questions For the Basic and Clinical Sciences. National Board of Medical Examiners.
- Dent, J.A.; Harden, R.M. (2005). A Practical Guide For Medical Teachers. Edinburgh: Elsevier -Churchill Livingstone.
- Schuwirth et al (2003). ABC of learning and teaching in medicine – Written assessment. In: BMJ, 2003; 326: 643-645.
- Schuwirth, W.T.; Van der Vleuten, C. P.M. (2004). Different written assessment methods: what can be said about their strengths and weaknesses?. In: Medical Education, 2004; 38: 974-979.