Avaliação Somativa de Habilidades Cognitivas: Experiência Envolvendo Boas Práticas para a Elaboração de Testes de Múltipla Escolha e a Composição de Exames

Summative Assessment of Cognitive Skills: an Experience Involving Good Practices for Writing Multiple Choice Tests and Exam Composition

Valdes Roberto Bollela[1]
Marcos de Carvalho Borges[1]
Luiz Ernesto de Almeida Troncon[1]

PALAVRAS-CHAVE

- Questões de Exames.
- Testes de Aptidão.
- Cognição.
- Avaliação Educacional.
- Docentes.
- Cursos de Treinamento.
- Educação Médica.

RESUMO

As questões objetivas ou testes de múltipla escolha com somente uma alternativa correta constituem um dos métodos mais utilizados em todo o mundo em exames destinados a avaliar habilidades cognitivas, especialmente nas avaliações somativas. Provas que contêm predominantemente questões objetivas de múltipla escolha são utilizadas sobretudo nos exames em que muita coisa está em jogo, como concursos vestibulares, provas finais de cursos de graduação e exames próprios dos concursos de ingresso à residência médica ou de obtenção de título de especialista. Esta ampla difusão justifica-se pelo fato de os exames compostos com este tipo de questão preencherem mais completamente os requisitos de validade e de fidedignidade, além de terem vantagens quanto à viabilidade ou factibilidade, particularmente em provas com grande número de candidatos. No entanto, os requisitos de validade e fidedignidade, em especial, somente são preenchidos adequadamente quando se seguem normas próprias das boas práticas de construção de exames e de elaboração dos testes propriamente ditos. Neste artigo se descrevem algumas das boas práticas na elaboração de testes de múltipla escolha, baseadas em fontes da literatura nacional e internacional, bem como na experiência dos autores. Apresenta-se e se discute um conjunto de regras práticas para construir questões de múltipla escolha de boa qualidade no que se refere à forma e ao conteúdo e se comenta como compor tabelas de especificação. Este tipo particular de matriz da avaliação permite verificar o alinhamento entre os temas abordados na prova e os objetivos curriculares ou o que se espera que os estudantes/candidatos dominem, o que configura um importante indicador de validade. Apresenta-se também uma experiência bem-sucedida de trabalho em grupo na organização de exames que utilizam este tipo de questão, como exemplo de desenvolvimento de processo organizado para obtenção de questões de melhor qualidade, que também contribuiu para o desenvolvimento docente no campo de avaliação da aprendizagem.

[1] *Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil.*

KEY-WORDS

– Examination Questions.

– Performance Tests.

– Cognition.

– Educational Measurement.

– Faculty.

– Training.

– Education, Medical.

ABSTRACT

Objective items or multiple choice questions with just one correct answer are among the most widely used methods for cognitive skills assessment, especially in exams designed for summative purposes. Assessments related to the cognitive domain using multiple choice questions are mostly used in high-stake exams, i.e. where the risks of failing are associated with serious consequences for the candidates. The widespread use of objective items for assessing learning in the cognitive domain may be explained by the fact that this exam modality fulfills both validity and reliability requirements, with the additional advantage that they are practical for use in exams with large numbers of candidates. Nevertheless, the validity and reliability requirements, in particular, will only be properly fulfilled when the process of writing multiple choice questions follows the rules of good practices for constructing exams and writing tests. This manuscript describes some of the rules for developing high quality multiple choice tests, based on both national and international published sources, as well as on the author's experience. These rules relate to the content, language and presentation of the questions. This paper also addresses the importance of following appropriate rules for blueprinting construction, in order to show the alignment between assessment and curriculum and thereby contribute to meeting the validity requirements. It also briefly describes and discusses a successful experience of team work for constructing items and organizing exams. This experience exemplifies the combination of an organized process for constructing high quality questions for a well-balanced examination with an institutional strategy for faculty development in the field of learning assessment.

Recebido em: 21/2/17

Aceito em: 27/4/17

INTRODUÇÃO

As questões de múltipla escolha (QME) constituem um dos métodos mais utilizados em todo o mundo em exames destinados a avaliar habilidades cognitivas. Este amplo uso é justificado pelas inúmeras vantagens que esta modalidade de avaliação possui, não obstante apresentar, também, desvantagens e limitações. É fundamental ressaltar que as inúmeras vantagens da utilização das QME somente estarão presentes e justificarão a sua utilização se forem seguidas normas próprias das boas práticas de construção de exames e de elaboração dos testes propriamente ditos.

Este artigo tem como objetivos descrever de modo sucinto algumas das boas práticas na elaboração de QME com apenas uma alternativa de resposta correta e apresentar uma experiência exitosa de trabalho em grupo na organização de exames para a residência médica utilizando este tipo de questão. Para o bom entendimento e a adequada contextualização destas práticas, abordam-se, inicialmente, princípios básicos da avaliação do estudante, com foco na avaliação de habilidades cognitivas na área da saúde.

PRINCÍPIOS BÁSICOS EM AVALIAÇÃO DE HABILIDADES

A avaliação do estudante ou do profissional em treinamento pode ser definida de forma mais restrita ou de um modo mais amplo. Na forma mais restrita, avaliar significa verificar se o estudante atingiu os objetivos instrucionais preestabelecidos, ou, em palavras mais simples, se aprendeu o que deveria ter aprendido[1]. Na definição mais ampla[2,3], o avaliar compreende processos de obtenção de informações sobre o desempenho do estudante em diferentes domínios, de modo a cumprir três funções principais: fomentar o aprendizado (avaliação formativa), embasar decisões que terão implicações em seu progresso (avaliação somativa) e contribuir para o controle da qualidade dos programas educacionais (avaliação informativa).

Aprender significa adquirir habilidades e desenvolver competências em diferentes domínios. A classificação mais tradicional das habilidades, conhecida como "Taxonomia de Bloom"[4], baseia-se no tipo de objetivos educacionais e inclui três domínios: cognitivo (conhecimento, raciocínio), psicomotor (procedimentos) e afetivo (valores, opiniões, juízos, atitudes). O domínio cognitivo, em sua proposição original, englobava seis níveis de complexidade crescente (Quadro 1). Uma proposta de modificação mais recente da proposição original[5]

http://dx.doi.org/10.1590/1981-52712015v42n4RB20160065

inclui seis processos mentais, que se aplicariam de modo distinto a quatro dimensões do domínio cognitivo[6].

Quadro 1
Níveis de complexidade e dimensões das habilidades cognitivas segundo a proposição original de Bloom e colaboradores[4] e de acordo com a modificação de Anderson e colaboradores[5]

Proposição Original	Proposição de Modificação	
Taxonomia de Bloom	Níveis de complexidade	Dimensões
Conhecimento	Relembrar	
Compreensão	Entender	Factual
Aplicação	Aplicar	Conceitual
Análise	Analisar	Procedimental
Síntese	Avaliar	Metacognitiva
Avaliação	Criar	

Na avaliação de habilidades referentes ao domínio cognitivo, temos simplificado estas classificações[7] utilizando três níveis referenciais para a elaboração de questões:

I. Básico – conhecimento factual, que envolve basicamente a memorização de fatos e sua recuperação (*recall*);
II. Intermediário, que envolve não só o conhecimento adquirido, mas as capacidades de compreensão, interpretação e aplicação daquele conhecimento para resolver problemas mais simples;
III. Avançado, que envolve os níveis de análise, síntese e avaliação para propor soluções para problemas mais complexos, como os que habitualmente se apresentam ao profissional da área da saúde.

No planejamento de qualquer procedimento de avaliação, convém inicialmente considerar as suas finalidades e ao mesmo tempo concentrar-se no "que" deverá ser avaliado, ou seja, quais os domínios, habilidades e competências que serão avaliados. Esta é uma etapa geralmente negligenciada e que pode colocar em risco todo o processo avaliativo.

A etapa subsequente envolve a escolha dos métodos que melhor se adaptem aos domínios a serem avaliados e às finalidades da avaliação. Nas avaliações somativas, devem ser consideradas todas as implicações que os resultados podem trazer para quem será avaliado. Na língua inglesa, estas implicações são denominadas *stakes*, gerando os termos *high-stakes* e *low-stakes*, que se aplicam especialmente a exames. O primeiro destes termos se aplica aos exames em que muita coisa está em jogo (por exemplo, provas finais de cursos de graduação, concursos vestibulares, de título de especialista ou de ingresso à residência médica). Os exames do tipo *low-stakes* são aqueles em que as implicações têm importância relativamente menor ou aquelas em que eventuais resultados adversos podem ser compensados de outras maneiras (por exemplo, as provas parciais de disciplinas de graduação).

A escolha do método a ser empregado deve, então, ser pautada pela finalidade da avaliação, pelas implicações de seus resultados (*high-stakes versus low-stakes*) e, em especial, pelos domínios aos quais se relacionam as habilidades e competências que constituem o foco da avaliação. Nesta escolha, é essencial considerar os atributos gerais dos métodos de avaliação[8], que são apresentados no Quadro 2.

Quadro 2
Atributos gerais dos métodos de avaliação

Atributo	Significado
Validade	O método permite avaliar de fato o que se pretende avaliar
Fidedignidade ou confiabilidade	O método tem precisão e é reprodutível
Viabilidade	O método pode ser aplicado com os recursos disponíveis
Aceitabilidade	O método é aceito por todos em função de suas qualidades
Equivalência	Não há diferenças significativas quando o método é aplicado ao mesmo tempo em diferentes locais
Impacto educacional	A avaliação pode afetar positiva ou negativamente educandos, professores e instituições.
Efeito catalítico	O emprego do método pode fomentar o desenvolvimento educacional da instituição

Tendo em vista o escopo deste artigo, abordaremos a seguir somente três destes atributos: validade, fidedignidade (confiabilidade) e impacto educacional.

A validade de uma avaliação estará garantida quando for possível demonstrar que o método escolhido avalia exatamente aquilo que se pretendia que o estudante aprendesse (alinhamento com objetivos instrucionais estabelecidos e competência esperada), o que indica que os resultados da sua aplicação podem ser generalizados. De modo geral, a validade depende da adequação do método à natureza do domínio que se pretende avaliar, bem como se é coerente com o processo formativo.

A validade também depende da apropriada amostragem de conteúdos e tarefas que serão incluídos na avaliação, ou seja, se incluirmos apenas uma questão sobre um tema no exame final de um curso que teve duração de seis meses e que

http://dx.doi.org/10.1590/1981-52712015v42n4RB20160065

abordou cerca de 20 temas, provavelmente a amostragem não terá sido adequada. Nas avaliações do domínio cognitivo, a validade é, portanto, determinada pelo número apropriado de questões, que deve ser proporcional à extensão dos conteúdos que devem ser avaliados. As questões devem abordar temas relevantes e representativos e exigir tarefas mentais e níveis de complexidade e dificuldade condizentes com o estágio de formação do educando. A redação e a forma de apresentação das questões podem afetar de modo importante a validade dos exames, sendo por isto essencial que a sua elaboração atenda às boas práticas. Além disso, as condições de aplicação do método escolhido na avaliação, em termos de tempo alocado para as tarefas, resposta às questões e ambiente físico, devem ser, também, adequadas.

A fidedignidade, ou confiabilidade constitui atributo que se relaciona à precisão, objetividade e consistência do método, características que, quando presentes, indicam que os resultados da sua aplicação são reprodutíveis. A fidedignidade é determinada pelo adequado controle das variáveis internas e externas capazes de influenciar os resultados do exame. Muitos dos fatores que influenciam a validade da avaliação podem também exercer efeito sobre a sua fidedignidade. Estes fatores são, de modo geral, relacionados a três aspectos: elaboração, aplicação e correção do exame. Assim sendo, quanto maior for o número de questões e o tempo de duração da prova, mais provável será que se atinjam níveis elevados de fidedignidade. Estes também dependem criticamente da igualdade de condições de aplicação do exame para todos os examinandos e do controle das influências externas que podem estar presentes quando da sua realização. Os critérios de correção devem estar estabelecidos de forma clara e ser utilizados de modo padronizado pelos avaliadores. Sempre que possível, deve-se cuidar para que a correção do exame se faça de forma isenta de influências externas.

É importante ter em conta que tanto a validade como a fidedignidade dos exames podem ser estimadas com técnicas específicas, que podem ser aplicadas antes ou após a realização dos procedimentos avaliativos. A apresentação destas técnicas e a discussão do seu significado fogem, porém, do escopo deste artigo.

O impacto educacional pode ser definido pelo potencial de um método ou procedimento avaliativo, pela sua própria existência ou, pelo modo como é empregado, influenciar positiva ou negativamente os processos de ensino e aprendizado. O impacto educacional positivo se caracteriza quando a própria existência de avaliações motiva e encoraja os estudantes. Estes podem, também, aprender mais e integrar conhecimentos durante a avaliação, dependendo do modo como o exame é construído e são elaboradas as questões. Em todo exame, mas principalmente naqueles empregados nas avaliações somativas do tipo *high-stakes*, a escolha dos conteúdos e a forma de elaboração das questões podem sinalizar para estudantes e professores o que é importante em termos de conteúdo e de processos mentais exigidos, influenciando, assim, o processo educativo daqueles que se submeterão ao mesmo tipo de exame no futuro.

Nas avaliações somativas do domínio cognitivo, a presença de evidências de validade e de fidedignidade nos exames atesta a qualidade dos procedimentos, sendo essencial, sobretudo, nas avaliações do tipo *high-stakes*. No entanto, investir para que os exames atinjam altos níveis de fidedignidade somente faz sentido se forem também preenchidos os critérios de validade. É importante, porém, ter em conta que a presença dos atributos de validade, fidedignidade e impacto educacional não constitui propriedade intrínseca aos métodos, mas é determinada pelo modo como eles são utilizados, o que, por sua vez, depende dos níveis de treinamento e de capacitação da equipe que elabora os processos avaliativos.

MÉTODOS DE AVALIAÇÃO DO DOMÍNIO COGNITIVO

O domínio cognitivo pode ser avaliado mediante o emprego de diferentes métodos, que fazem parte do elenco mais tradicional de recursos da avaliação educacional[9]. Este conjunto inclui as modalidades apresentadas no Quadro 3.

Quadro 3
Métodos para avaliação do domínio cognitivo[9]

- Dissertação (ensaio)
- Ensaio curto
- Questões de resposta aberta longa
- Questões de resposta aberta curta
- Questões estruturadas (substituição, preenchimento de lacunas).
- Testes objetivos
- Provas orais
- Portfólio
- Autoavaliação

Entre estes métodos de avaliação do domínio cognitivo, as QME, também denominadas testes objetivos ou questões da melhor resposta única (*single best answer* ou *one best answer*), ganharam proeminência sobretudo em exames do tipo *high-stakes*, em função de apresentarem inúmeras vantagens, que superam algumas desvantagens e limitações que certamente apresentam. As QME podem também ser elaboradas em diversos formatos[10,11], como mostra o Quadro 4.

http://dx.doi.org/10.1590/1981-52712015v42n4RB20160065

Quadro 4
Modalidades de testes objetivos para avaliação do domínio cognitivo[10]

- Afirmações do tipo falso ou verdadeiro
- Resposta múltipla (múltiplo "falso ou verdadeiro")
- Completar afirmações
- Completar lacunas
- Associação
- Asserção – razão
- Interpretação de textos, dados, tabelas ou gráficos
- Item de resposta única (múltiplas alternativas e única resposta)

VANTAGENS E LIMITAÇÕES DOS EXAMES COM TESTES OBJETIVOS

Entre as principais vantagens das QME estão as possibilidades de preencherem mais completamente os requisitos de validade e de atingirem níveis mais elevados de fidedignidade. As QME permitem também realizar a avaliação do domínio cognitivo nas mais variadas áreas, em seus vários níveis de complexidade. Com elas é possível construir exames que contemplem amostragem adequada do conteúdo a ser avaliado, utilizando grande número de questões, como, por exemplo, os exames vestibulares e de acesso à residência médica. Estas podem ser elaboradas de modo a apresentar situações, problemas e condições análogos aos observados na prática profissional. Estas vantagens são especialmente aplicadas ao formato de teste objetivo denominado "item de resposta única"[11], que apresenta maior facilidade para representar problemas e situações reais. É importante salientar que as boas práticas de elaboração de questões, que serão apresentadas e discutidas a seguir, encontram-se muito bem estabelecidas para este tipo específico de formato.

Adicionalmente, quando se empregam QME, é possível fazer o controle adequado das condições de realização do exame, mesmo quando aplicados a grande número de candidatos e em diferentes locais. Além destas vantagens, sobressai a facilidade de correção, seja na forma manual ou de modo automatizado, já que não há necessidade de composição de critérios de correção mais elaborados, mas somente indicar a alternativa correta para cada questão. Adicionalmente, constitui importante vantagem a facilidade para comunicar os resultados (gabarito) e com isso fornecer uma devolutiva aos examinandos seja sobre o seu desempenho geral, seja sobre o que se refere a áreas específicas de conteúdo avaliadas.

Após a aplicação dos exames com QME, um procedimento técnico denominado "análise de itens" possibilita avaliar tanto a qualidade do exame como um todo, como a de cada questão, fornecendo dados sobre a consistência do exame e a variabilidade das respostas que permitem construir índices de fidedignidade. Além disso, este tipo de análise permite verificar o "desempenho" de cada questão em termos de alternativas mais escolhidas, de dificuldade geral e de discriminação entre os candidatos que tiveram maior e menor aproveitamento no exame como um todo, o que vai auxiliar nas conclusões sobre a validade do processo avaliativo. Os dados desta análise devem ser apresentados aos organizadores do exame e elaboradores de questões, pois constituem oportunidades de aprendizagem e desenvolvimento para os professores, o que resulta em importante impacto educacional positivo do processo como um todo.

Como já mencionado, os exames com QME não são livres de desvantagens e limitações, que devem ser consideradas quando se opta por este método de avaliação do domínio cognitivo. A primeira delas é que é impossível utilizar este recurso para a avaliação isolada e exclusiva de certas habilidades, como a do raciocínio profundo específico para certos tópicos em determinadas áreas. A necessidade de delimitação estreita e muito precisa do aspecto que vai ser abordado em cada questão pode dificultar a avaliação de situações com múltiplos componentes. Além disso, nas profissões da saúde abundam problemas e situações para os quais não existe uma única solução correta ou inexiste o consenso bem firmado sobre o que é mais adequado. Isto torna as QME inadequadas para a avaliação cognitiva destes problemas e situações.

Desde a introdução das QME no elenco de recursos da avaliação educacional, conhece-se bem a limitação representada pela possibilidade de acerto casual ("chute" ou *guessing*), que é tanto maior quanto menor o número de alternativas de respostas. Ademais, sempre é preciso ressaltar a precária validade do emprego das QME para a avaliação de outros domínios que não o cognitivo. Por fim, é importante considerar que a prática sistemática e exclusiva da avaliação com o emprego de QME pode trazer forte impacto educacional negativo por induzir o direcionamento do tempo e da energia dos estudantes e candidatos, bem como dos professores, em algumas circunstâncias, no treinamento para responder a questões e não no aprendizado propriamente dito.

É importante salientar, porém, que as aludidas vantagens das QME sobre outros formatos, especialmente as que se referem à possibilidade de preencherem mais completamente os requisitos de validade e de fidedignidade, ocorrem somente se os exames forem bem planejados e as questões bem elaboradas, seguindo as recomendações apresentadas a seguir.

BOAS PRÁTICAS NA ELABORAÇÃO DE QUESTÕES

A seguir, apresentaremos de forma sucinta as principais recomendações para a elaboração de QME do tipo *one best answer*, buscando ilustrar cada uma delas. Apresentaremos também

http://dx.doi.org/10.1590/1981-52712015v42n4RB20160065

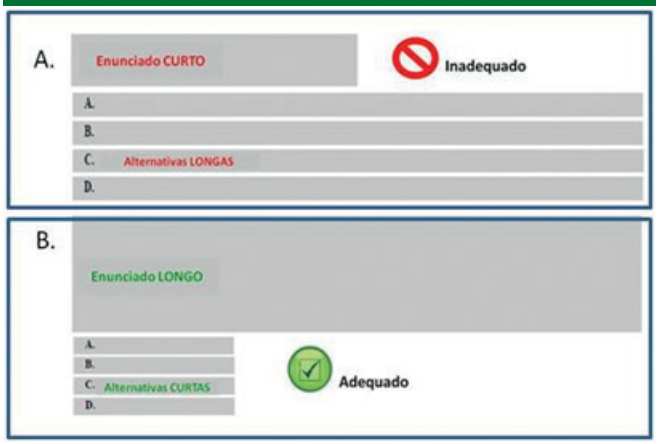
uma proposta de estrutura para QME utilizadas nas avaliações cognitivas mais comuns na área da saúde, como o teste de progresso e os concursos para vagas de residência médica e títulos de especialista. Finalmente, comentaremos sobre a matriz de especificação de conteúdo da avaliação, ou o *blueprinting* da prova, que é um componente fundamental para a definição da validade da prova[7,10-14].

REGRAS BÁSICAS E ESSENCIAIS NA ELABORAÇÃO DE QME

1. Cada teste deve focar um conceito importante: tendo como referência os objetivos de aprendizagem e as competências a serem adquiridas, devem-se privilegiar questões que abordem situações clínicas relevantes e prevalentes ou situações clínicas potencialmente catastróficas, em que o conhecimento avaliado é fundamental para a correta interpretação ou condução do caso. Recomenda-se que o elaborador do teste não desperdice tempo com questões que apenas avaliem conhecimento sobre fatos triviais e sem importância. Deve-se dar preferência e focar em fatos que são habitualmente encontrados na vida real. Além disso, é fundamental evitar questões muito complexas ou "pegadinhas".

2. Estrutura do enunciado versus alternativas: de modo geral, o enunciado de um teste deve ser maior (mais longo) que o conjunto das alternativas. Esta é uma regra muito importante e pode ser facilmente checada apenas observando-se o "desenho geral do teste". A Figura 1 apresenta o aspecto geral de um teste considerado "adequado" quando comparado com um teste considerado "inadequado".

Figura 1
Aspecto geral de um teste considerado "adequado" quando comparado a um teste considerado "inadequado"



Na Figura 1A, observamos um teste com desenho geral inadequado, onde o enunciado é bem curto e as alternativas longas, o que deve ser evitado. Na Figura 1B, temos o enunciado longo e as alternativas curtas, o que é considerado adequado.

3. Cada QME deve avaliar a aplicação do conhecimento e não apenas a lembrança de fatos isolados: neste caso, o uso de vinhetas (breves casos clínicos) deve constituir a base para um bom enunciado. As vinhetas devem conter apenas parte das informações clínicas, e não todos os dados clínicos existentes. No entanto, as informações devem sempre ser apresentadas numa ordem específica. Inicia-se normalmente com os dados do paciente e a sua principal queixa. Em seguida, apresentam-se os dados da história e os achados do exame físico (sintomas e sinais) e, sempre que necessário, os resultados de exames complementares, o tratamento inicial e a evolução do quadro, a depender do foco que a questão terá.

4. O enunciado deve definir sempre uma situação concreta e conter no seu final uma questão bem clara a ser respondida: para checar se a questão está adequada, devem-se cobrir as alternativas e ver se é possível respondê-la apenas com as informações do enunciado. Quando isto ocorre, indica um bom enunciado; caso contrário, a questão deve ser revista, com a reconstrução da vinheta clínica, ou da pergunta do teste, ou das alternativas.

5. Todos os distratores devem ser homogêneos e acompanhar a alternativa correta: se um teste é elaborado com vistas a perguntar sobre a melhor conduta para determinada situação clínica e a resposta correta for um medicamento, como, por exemplo, a colchicina, seria desejável que todos os distratores (as demais alternativas não corretas) fossem também medicamentos de nomes curtos como o da colchicina. Assim, os distratores devem sempre pertencer à mesma categoria que a da resposta correta (diagnósticos; exames complementares; prognóstico, tratamento). Outra recomendação é evitar o uso de respostas compostas, em que cada alternativa contenha duas ou três respostas e não apenas uma. Recomenda-se, portanto, evitar situações como: a conduta é fazer "ISTO" e "AQUILO", ou qual o DIAGNÓSTICO e a CONDUTA, ou deve-se fazer "ISSO" por causa "DAQUILO". Esta situação, que é bastante comum quando ainda não se domina a prática de elaboração, é exemplificada no teste a seguir:

Questão: Qual alternativa corresponde ao tratamento mais efetivo para a crise aguda de gota?

a) calor local e diclofenaco

b) aspirina e gelo
c) alopurinol e colchicina
d) colchicina e prednisona
e) diclofenaco e colchicina → resposta correta

Nota-se neste teste que não há vinheta clínica (problema sério, o que contraria a recomendação número três) e, além disso, as alternativas de respostas são todas compostas. Observe que a palavra "colchicina" repete-se por três vezes, e o medicamento "diclofenaco" repete-se por duas vezes, indicando, pela repetição, que a alternativa que os contém pode ser a resposta correta. Outro problema deste teste é a existência de distratores de diferentes categorias: enquanto uns são medicamentos, outros não o são (gelo, calor local), o que também deve ser evitado. Neste caso, tente reescrever os distratores buscando deixá-los mais homogêneos e focando em um único aspecto, evitando respostas compostas.

6. Os distratores devem ser coerentes com enunciado e a pergunta do teste: todos os distratores devem ser plausíveis, gramaticalmente consistentes, compatíveis logicamente e do mesmo tamanho da resposta correta. Quando, em vez de elaborarmos uma pergunta, fazemos uma questão do tipo "preencha a lacuna", estes erros costumam ocorrer com maior frequência. Esta situação é exemplificada no teste a seguir:

Questão: A hiperemia conjuntival, injeção ciliar, edema de córnea, pupila em médio midríase paralítica e dor ocular com sintomas neurovegetativos, como náuseas e cefaléia, são sinais e sintomas da:
a) uveíte
b) glaucoma agudo → resposta correta
c) conjuntivite
d) esclerite posterior

Veja que, da forma como foi escrito, este teste induz ao erro, já que a resposta correta é "glaucoma", um substantivo masculino, e o teste apresenta os sinais e sintomas de algo que deveria ser um substantivo feminino ("da"...). Além disso, neste teste não há uma vinheta ou problema clínico, mas, sim, uma questão do tipo *recall*, que requer apenas memorização, o que deve também ser evitado (de novo, a recomendação número três).

7. Evite erros técnicos primários: evite escrever uma questão do tipo verdadeiro/falso (V ou F) no formato de um teste de múltipla escolha. A forma mais comum de se fazer isso é incluir no enunciado frases como: qual das seguintes alternativas é correta?; todas as afirmações seguintes estão corretas, EXCETO; marque a alternativa INCORRETA. Se a proposta for elaborar testes V ou F, o melhor seria seguir as recomendações para a elaboração deste tipo específico de questão objetiva, mas para QME do tipo *one best answer* estas frases nunca devem ser utilizadas.

8. Uso de histórias completas de pacientes reais: em geral, recomenda-se não utilizar na íntegra a história de um paciente real (aquele que está internado na enfermaria ou em atendimento no ambulatório) como a vinheta de testes de múltipla escolha, especialmente para avaliar estudantes. Habitualmente, a situação clínica de pacientes reais costuma ser muito complexa, de modo a gerar dúvidas. Os aspectos às vezes confusos dos casos reais inerentes à prática clínica nem sempre são aquilo que se pretende avaliar com testes de múltipla escolha. É possível se inspirar em casos reais, mas deve-se desenhar o teste tendo em vista o conhecimento e a sua aplicação que se pretende avaliar com a questão, que devem ser o foco do teste. Histórias complexas e pouco claras podem levar o estudante para longe da resposta que esperamos que ele encontre, e isto não é desejável.

9. Uso de valores de referências de exames na avaliação: é bastante apropriado e recomendado prover os estudantes com informações que estão habitualmente disponíveis na prática profissional cotidiana, como, por exemplo, os valores de referência de normalidade para exames complementares. Eles podem ser apresentados no início do caderno de provas ou junto a cada questão. Lembre-se de que as questões devem avaliar o conhecimento e sua aplicação a situações clínicas relevantes e não a memorização de coisas, como valores de referência de exames complementares, que, além disso, podem variar dependendo da técnica empregada em cada centro.

10. Uso de siglas, linguagem coloquial e dados não verdadeiros ou ambíguos: deve-se evitar ao máximo utilizar siglas, mas, se o fizer, deixe sempre explícito o seu significado. Embora algumas siglas sejam de conhecimento geral, muitas variam de um local para outro, o que pode gerar problemas de entendimento. Evite, também, utilizar a linguagem coloquial, ou seja, aquela que ouvimos do próprio paciente (a história contada com as palavras do paciente), inclusive porque há diferenças regionais de linguagem, e isto pode representar uma dificuldade adicional para estudantes/candidatos que não as conhecem. Apesar de ser frequente na vida real, jamais deve ser criada uma vinheta que contenha uma história clínica em que as informações do paciente não correspondem à realida-

http://dx.doi.org/10.1590/1981-52712015v42n4RB20160065

de. Por exemplo, um paciente que afirma que nunca teve relação sexual desprotegida e a questão deste caso é sobre doenças sexualmente transmissíveis, tendo como resposta correta a alternativa "gonorreia". Nas questões de múltipla escolha, o paciente deve sempre falar a verdade, ou a interpretação do médico deve ser oferecida no enunciado. Evite escrever frases ambíguas como: "O paciente diz que toma apenas uma dose de cachaça por dia", mas apresenta sinais de embriaguez.

11. Submeta o seu teste ao "teste da gaveta": após finalizar a primeira versão do seu teste, deixe-o guardado por alguns dias e depois o retome e leia novamente. É bastante comum perceber algumas inconsistências que precisam ser corrigidas e ajustadas. Quando existe um grupo de elaboradores trabalhando juntos, esta revisão pode ser feita na forma de um painel, onde cada questão é projetada e lida, e os presentes podem sugerir adequações e ajustes ao autor do teste.

Apesar de muito simples, este conjunto de 11 regras e recomendações básicas é muito efetivo e eficiente para garantir a qualidade na elaboração de testes de múltipla escolha. Seu emprego deve resultar em exames de melhor qualidade, levando a melhores conclusões sobre o aprendizado ou decisões mais corretas sobre os candidatos mais qualificados em processos seletivos classificatórios.

PROPOSTA DE ESTRUTURA DE TESTES DE MÚLTIPLA ESCOLHA PARA EXAMES NA ÁREA DA SAÚDE

Apresenta-se a seguir uma proposta de estrutura padronizada para elaboração de QME para exames somativos, que foi adaptada de uma fonte local[7] e de uma referência bastante tradicional[11]. Os componentes desta proposta são apresentados na Figura 2. Nela, o elaborador registra dados e características do teste que elaborou, informando inicialmente a sua área de origem (Cirurgia, Pediatria, etc.), a especialidade, quando for o caso (Urologia, Proctologia, Neurocirurgia, etc.) e o conteúdo do teste (manejo de hemorragia digestiva alta, diagnóstico de nódulo pulmonar, etc.). Em seguida, escreve o enunciado e a pergunta do teste e as alternativas de resposta, deixando claro qual é a correta e a comentando brevemente. É desejável que se comente também cada um dos distratores, justificando por que não são as respostas corretas. Por fim, recomenda-se oferecer pelo menos uma referência bibliográfica (livro, diretrizes, consensos) que suporte as explicações dadas. Esta padronização ajuda a qualificar a elaboração de testes em concursos ou exames muito grandes e com muitos elaboradores, além de ser extremamente útil na eventualidade de haver recursos contra a prova.

Figura 2
Estrutura de padronização da elaboração de questões de múltipla escolha, especialmente útil para concursos de residência médica

1. ÁREA DE ORIGEM: → Especialidade (se for o caso):
2. CONTEÚDO:
3. ENUNCIADO
   a. Vinheta clínica:
   b. Questão:
4. ALTERNATIVAS
   a. .
   b. .
   c. .
   d. .
5. Informar a ALTERNATIVA CORRETA
6. Comentários sobre a ALTERNATIVA CORRETA
7. Comentários sobre os DISTRATORES:
8. REFERÊNCIA(S) BIBLIOGRÁFICA(S):

Outros tipos de dados podem ser acrescentados a este modelo, dependendo do interesse e propósito da avaliação, tais como descrever o local onde o atendimento descrito no enunciado do teste está sendo realizado: atenção básica, pronto atendimento, enfermaria de um hospital geral, ambulatório especializado, terapia intensiva, etc. Cada teste pode também ser caracterizado em relação à área médica e especialidade da condição apresentada, tipo de condição clínica e foco principal do conteúdo, ou seja, se o teste é sobre diagnóstico, fisiopatologia, manejo clínico ou promoção da saúde. Esta classificação é de suma importância, pois com ela é possível construir a tabela de especificações da prova ou o *blueprinting*, que nada mais é que a matriz que alinha os temas abordados na prova com aqueles que se espera que os estudantes/candidatos conheçam e dominem para a prática profissional competente. É altamente desejável que isto ocorra, pois este é mais um indicador de validade da própria avaliação[13,14].

BOAS PRÁTICAS EM AÇÃO

Relata-se a seguir a experiência de um grupo de elaboradores de QME para o concurso de residência médica do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (HC-FMRP-USP), que organiza anualmente um exame de conhecimentos para cerca de 2.500 candidatos às vagas nos programas de Clínica Médica. Em 2011, esta comissão, com o aval da chefia do Departamento de Clínica Médica e da Comissão de Residência Médica (Coreme) do HC-FMRP-USP, mudou a forma de elaborar e revisar as QME para os concursos anuais de seleção da instituição. Em 2013, uma equipe da Clínica Médica começou a auxiliar também os outros programas na elaboração e revisão de testes.

A elaboração de QME de boa qualidade, seguindo as boas práticas e tendo impacto educacional positivo, requer treino e bastante prática. Habitualmente, quando solicitados a produzirem questões para um exame, os elaboradores criam novos testes, trabalhando individualmente, e os encaminham a um coordenador de área, que seleciona aqueles que julga serem os mais adequados. Apesar de amplamente utilizada, esta prática vinha se mostrando pouco efetiva e apresentava inúmeras desvantagens, que poderiam prejudicar a validade e a fidedignidade da prova.

Quando um teste é elaborado individualmente e não passa por um processo de revisão em grupo (por pares), ele fica sujeito a apresentar diversas impropriedades de forma e conteúdo. Habitualmente, as questões são elaboradas por especialistas em determinada área que têm tendência natural a achar que o tema da questão é essencial e presumem que todos os que se submeterão ao exame deveriam ter aquele conhecimento. Desta maneira, muitas vezes, o teste não fica adequado ao nível do estudante que está sendo selecionado para o primeiro ano de um programa de residência médica, o que pode prejudicar a validade da prova. Além disso, a elaboração individual não garante a homogeneidade da prova como um todo, uma vez que as questões podem estar em formatos distintos, nem sempre adequados. Outra inadequação que pode ocorrer com a elaboração individual da questão é que a resposta correta pode estar baseada somente na opinião individual do elaborador ou nas práticas da instituição, e não necessariamente em consensos nacionais ou internacionais. Esta ocorrência é comum e com freqüência gera recursos que, quando atendidos, podem demandar mudanças de gabarito. Por outro lado, quando se faz a revisão das questões por um painel de vários elaboradores, os ajustes de forma e a verificação da pertinência de conteúdos, assim como a adequação das respostas corretas, são feitos com mais propriedade.

A elaboração individual das QME restringe também a criação de questões multidisciplinares, que podem refletir com mais precisão a realidade. Outro aspecto que pode melhorar com a revisão feita pelo painel de elaboradores é a maior facilidade para definição de distratores homogêneos, considerando que todos devem ser plausíveis, gramaticalmente consistentes, compatíveis logicamente e do mesmo tamanho da alternativa correspondente à resposta correta.

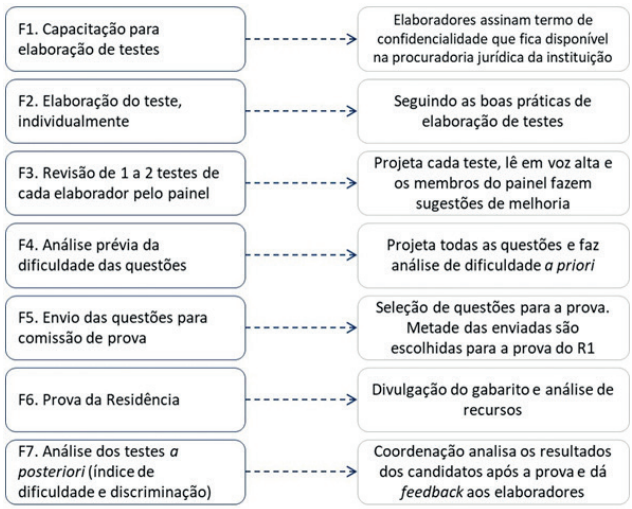
Entre as inconformidades mais frequentemente observadas, podemos citar: QME escritas no formato do tipo "verdadeiro ou falso"; questões que não avaliam o raciocínio, mas apenas a memorização de fatos ou tópicos isolados; questões sem vinheta clínica no enunciado; ausência de pergunta clara no enunciado do teste; erros gramaticais e de concordância; questões que abordam conceitos ou condutas não consensuais na literatura ou prática médica; questões com respostas compostas ou que avaliam mais de uma categoria/domínio, etc. Em especial, esse tipo de inadequação, inevitavelmente, aumenta o número de questionamentos sobre alguns testes da prova e, portanto, o número de recursos impetrados pelos candidatos com o intuito de cancelar a questão e/ou alterar seu gabarito.

Em 2011, a comissão de provas do Departamento de Clínica Médica da FMRP-USP iniciou uma nova prática na elaboração e revisão de testes para o concurso da residência médica, que incluiu dois encontros de capacitação dos elaboradores em "Aspectos gerais da avaliação do estudante", com ênfase especial na avaliação somativa do domínio cognitivo, do tipo *high-stake*, que é a principal característica do exame para os programas de residência médica do HC-FMRP-USP. Além da capacitação de um grupo de dez elaboradores e de dois coordenadores de prova, foram promovidos também oito encontros para leitura e revisão conjunta das questões de Clínica Médica, que totalizaram 100 QME que seriam produzidas naquele ano. Este total incluiu 20 testes para a prova do acesso direto, 70 para a prova de acesso às especialidades e dez testes extras, que compuseram uma "reserva de segurança". Assim, cada elaborador preparou dez QME em um período de três meses, em que houve encontros para apresentação e revisão das questões pelo painel de elaboradores. A Figura 3 apresenta as diferentes fases deste trabalho, com descrição sucinta de cada uma delas.

A revisão em grupo é etapa fundamental na qualificação das QME, pois é a oportunidade de checar se as regras básicas foram seguidas, se o tema é relevante e se a questão está bem estruturada, buscando avaliar um aspecto relevante da prática profissional. Porém, em um concurso deste tipo, é fundamental tomar todas as providências para garantir a confidencia-

http://dx.doi.org/10.1590/1981-52712015v42n4RB20160065

Figura 3
Fases do trabalho de elaboração e revisão de itens (testes de múltipla escolha) no Departamento de Clínica Médica da FMRP-USP, visando às provas do concurso da residência médica

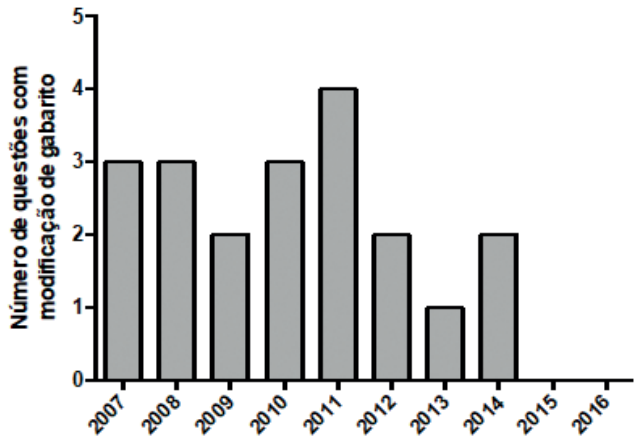


lidade e a segurança que a prova requer. Neste sentido, já na fase de capacitação (primeiro encontro), antes do início dos trabalhos de revisão, todos os participantes devem assinar um termo de sigilo e confidencialidade em relação ao conteúdo da prova, que fica arquivado na Coreme e procuradoria jurídica do HC-FMRP-USP. Este termo é renovado a cada edição do concurso.

Cada encontro ou reunião de trabalho tem duração de três horas, período em que os testes são revisados e modificados até que se atinja um formato adequado para a prova. Durante as revisões, verificam-se especificamente a pertinência do tema (validade), a redação e o formato de cada questão e se todas as recomendações de boas práticas que resultam em uma questão adequada foram atendidas. Verifica-se ainda se a resposta correta está devidamente embasada na literatura especializada. Nesse sentido, a escolha dos integrantes é fundamental para o bom andamento dos trabalhos do grupo, ou seja, é necessário escolher pessoas que tenham não somente disponibilidade, mas também conhecimento e perfil para o trabalho em equipe, situação em que são frequentes comentários e críticas sobre a produção de cada um. É necessário que haja o entendimento por parte dos elaboradores de que isto é parte do processo de trabalho e contribui para a qualidade do produto. Na última reunião de trabalho do grupo é feita uma análise *a priori* do grau de dificuldade de cada teste utilizando-se o método de Angoff modificado[15] e que serve de base para a escolha das questões pelo coordenador da prova, na tentativa de equilibrar a prova quanto ao grau de dificuldade, com proporções apropriadas de questões fáceis, moderadas e difíceis.

Nos últimos cinco anos, este conjunto de procedimentos tem sido utilizado com sucesso, visto que tem sido possível observar que tanto a qualidade das questões como todo o processo de seleção dos residentes têm ficado cada vez mais consistentes e livres de problemas. Um indicador indireto que reforça esta impressão é a redução significativa de recursos dos candidatos contra questões da prova ao longo dos últimos anos. A Figura 4 mostra o número de modificações do gabarito motivadas por recursos nas questões das provas de 2007 a 2015, em todas as cem questões das cinco áreas básicas (Ginecologia e Obstetrícia, Pediatria, Cirurgia, Medicina Social e Clínica Médica) da prova de acesso direto. É possível observar que este número vem diminuindo progressivamente na prova como um todo, e desde 2011 não houve mais nenhuma questão que tenha sido anulada ou sofrido mudança de gabarito nas 20 questões da área de Clínica Médica desta prova.

Figura 4
Número de questões que tiveram o gabarito modificado nas provas de seleção ao acesso direto no concurso para a residência médica do HC-FMRP-USP (divulgação autorizada pela Comissão de Residência Médica da instituição)



De modo geral, as outras áreas envolvidas na prova de acesso direto, além da Clínica Médica, têm aderido também à capacitação de elaboradores e a processos mais estruturados de revisão dos testes, sendo possível que isto esteja se refletindo também na melhoria deste indicador para o conjunto da prova nos últimos anos.

http://dx.doi.org/10.1590/1981-52712015v42n4RB20160065

Outro ponto importante na elaboração de exames, já comentado brevemente na segunda parte deste texto, é a construção do *blueprinting* da prova[10]. Trata-se da tabela de especificação da prova[10,13,14], que é uma matriz que procura alinhar todos os temas e as habilidades que serão avaliados com as questões incluídas na prova. Espera-se que os temas sejam relevantes e compatíveis com o que os estudantes ou candidatos devam conhecer e dominar, por serem relevantes para a prática profissional competente.

A elaboração do *blueprinting* é extremamente importante, sendo que uma matriz bem construída constitui mais um indicador de validade da própria avaliação. Geralmente ele é apresentado sob a forma de uma tabela em que constam informações como tema da questão (tuberculose, doença isquêmica cardíaca, etc.); foco da questão (fisiopatologia, diagnóstico, manejo); área envolvida (Reumatologia, Cardiologia, Pneumologia, etc.), cenário (unidade básica de saúde, pronto atendimento, hospital secundário, ambulatórios) onde o "paciente" da questão está sendo assistido, nível de complexidade do atendimento (atenção básica, nível secundário ou terciário). É possível também incluir dados sobre os índices psicométricos de cada teste (índice de dificuldade e discriminação de cada questão) quando eles tiverem sido calculados após a realização da prova. A análise do *blueprinting* de cada prova oferece a oportunidade de checar a diversidade dos temas e cenários e auxilia nas provas subsequentes, a fim de se evitar repetição de questões similares.

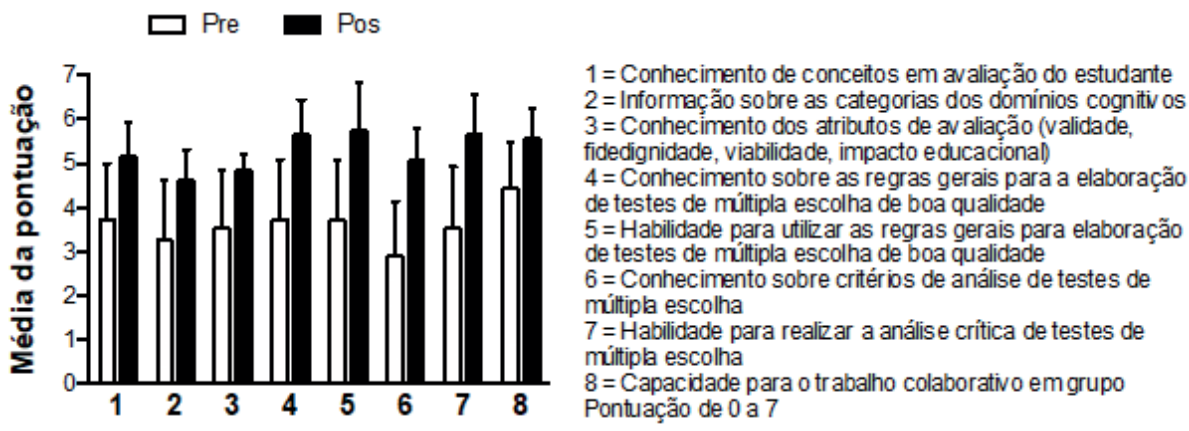
O grau de satisfação dos elaboradores em participar de um processo com este formato, bem como o seu ganho em conhecimentos e habilidades para elaborar QME foram avaliados com questionários específicos[16] nos anos de 2011 e 2012[17]. Os resultados mostraram que a maioria absoluta deles referiu estar satisfeita e indicou substancial ganho de conhecimentos e habilidades, expresso na comparação das percepções antes e após terem participado do processo.

Estes resultados indicam que iniciativas deste tipo permitem aliar a qualificação do processo de elaboração de testes a uma estratégia institucional de desenvolvimento docente no campo de avaliação da aprendizagem. Outro indicador de sucesso é o fato de que os integrantes do grupo, sistematicamente, manifestam interesse em participar do grupo de elaboradores ano após ano[17].

CONSIDERAÇÕES FINAIS

As QME com somente uma alternativa correta têm utilização muito difundida em todo o mundo na avaliação somativa de habilidades cognitivas, sobretudo na área da saúde. Ainda que esta modalidade de avaliação apresente limitações, estas são sobrepujadas pelas inúmeras vantagens que possui, o que permite preencher mais completamente os necessários requisitos de validade e de fidedignidade. No entanto, para que isto ocorra, é preciso atentar para as boas práticas de confecção de questões e de construção de exames, que envolvem recomendações relativas ao conteúdo e à forma dos testes, bem como à composição da matriz de avaliação ou tabela de especificação (*blueprinting*). Experiências de trabalho em grupo para elaboração de questões são efetivas para obter produtos de melhor qualidade e constituem também interessante estratégia institucional de desenvolvimento docente no campo de avaliação da aprendizagem.

Figura 5
Autopercepção dos participantes sobre conhecimento e domínio de habilidades para elaborar questões objetivas, antes e após o treinamento e o processo de elaboração/revisão de testes pelo painel de elaboradores



REFERÊNCIAS

1. Popham WJ. Educational evaluation, Englewood Cliffs (New Jersey), Prentice Hall, 1975.
2. Vianna HM. Introdução à avaliação educacional. São Paulo: IBRASA, 1989.
3. Troncon LEA. Estruturação de Sistemas para Avaliação Programática do Estudante de Medicina. Revista Brasileira de Educação Médica 03/2016; 40 (1):30-42. DOI:10.1590/1981-52712015v40n1e01392015
4. Bloom BSS, Englehart MD,Furst EJ, Hill, WH; Klathwohl,DR. Taxonomia de objetivos educacionais. Porto Alegre: Globo, 1976.
5. Anderson LA, Krathwohl D, Airasian P, Cruikshank KA, Mayer RE, Pintrich P, Raths J, Wittrock MC. A taxonomy for learning, teaching, and assessing: a revision of Bloom's Taxonomy of Educational Objectives. New York: Addison, Wesley Longman, 2001.
6. Panúncio-Pinto MP, Troncon LEA. Avaliação do estudante – aspectos gerais. Medicina (Ribeirão Preto) 2014;47(3):314-23.
7. Cianflone ARL, Troncon LEA, Rodrigues MLV, Figueiredo JFC. Recomendações para a elaboração de testes de múltipla escolha. Grupo de trabalho para avaliação, Comissão de Graduação, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 1994.
8. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. MedTeach. 2011;33(3):206-14. doi: 10.3109/0142159X.2011.551559.
9. Newble D, Cannon R. A Handbook for Medical Teachers. 4th Edition. Kluwer Academic Publishers. New York, Boston, Dordrecht, London, Moscow. 2002.
10. Vianna, H.M. – Testes em Educação. São Paulo, Ibrasa, 1983.
11. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. 3rd Edition. National Board of Medical Examiners, Philadelphia, PA, USA. 2002. (disponível em: *http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf*).
12. Jolly B. Written examinations, pages 208-231. In: Swanwick T. Undestanding Medical Educaton: Evidence, Theory and Practice. 1stEdition. Wiley-Blackwell. London
13. Coderre S, Woloschuk W, McLaughlin K. 2009. Twelve tips for blueprinting. Med Teach 31:359–361.
14. Bridge P, Musial J, Frank R, Roe T, Sawilowsky S. 2003. Measurement practices: Methods for developing content-valid student examinations. MedTeach 25(4):414–421.
15. Cusimano MD. Standard setting in medical education. Acad Med. 1996;71(10 Suppl):S112-20.
16. Bhanji F, Gottesman R, Grave W, Steinert Y, Winer LR. The Retrospective Pre–Post: A Practical Method to Evaluate Learning from an Educational Program. AcadEmerg Med. 2012; 19:189-194. doi: 10.1111/j.1553-2712.2011.01270.x
17. Borges MC, Elias PCL, Fernandes F, Costa NK, Oliveira RDR, Bollela VR. Continuing professional development on item writing: piggybacking on residency demands. In: Europe AfMEi, editor. AMEE 2012; Lyon 2012. p. 52.

APOIO FINANCEIRO

Fundação de Apoio ao Ensino, Pesquisa e Assistência do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (Faepa).

CONTRIBUIÇÃO DOS AUTORES

Valdes Roberto Bollela e Marcos de Carvalho Borges participaram da coleta dos resultados da experiência descrita; os três autores participaram do planejamento da experiência descrita e tiveram igual envolvimento na redação e na revisão final do manuscrito.

CONFLITO DE INTERESSES

Nenhum.

ENDEREÇO PARA CORRESPONDÊNCIA

Luiz E. A. Troncon
Hospital das Clínicas da FMRP
Departamento de Clínica Médica – Campus da USP
14049-900 – Ribeirão Preto – SP
E-mail: ledatron@fmrp.usp.br