RELATÓRIO DA COMISSÃO TÉCNICA PROVA TEÓRICA DO EXAME FINAL DO INTERNATO DE MEDICINA GERAL E FAMILIAR – ÉPOCA ESPECIAL DE SETEMBRO / OUTUBRO DE 2018

Para: Coordenações do Internato de MGF, ACSS/CNIM, Colégio de Medicina Geral e Familiar

Relatores: Isabel Santos, Bruno Heleno

Da elaboração do teste à sua aplicação

Terminada a época especial do Exame Final do IMMGF e serem conhecidos os respectivos resultados provisórios e definitivos da prova teórica, a Comissão Técnica de elaboração da Prova Teórica apresenta o seguinte relatório:
No dia 16 de Outubro de 2018 efectuaram a Prova Teórica 83 candidatos.
O lançamento das respostas dos candidatos foi efetuado numa plataforma electrónica que permite a obtenção imediata da classificação gerando uma folha das respostas lançadas, por cada candidato, em pdf.
Foi entregue aos candidatos, pelos júris, uma cópia da sua folha de respostas.
A chave do teste foi dada a conhecer aos candidatos no dia 16 de Outubro.
No dia 17 de Outubro, as Coordenações de Internato providenciaram a consulta dos testes, aos candidatos inscritos para esse efeito.
Disponibilizou-se uma ligação electrónica para envio de contestações.
Foram contestadas 40 perguntas do teste.

A distribuição das contestações superiores a 10 por pergunta

Pergunta	Contestações
90	11
76	12
6	14
25	14
32	15
48	16
86	17
85	18
17	24
45	32
55	37

As perguntas com menos de 20 % de respostas certas foram:

Perguntas	% de respostas certas
25	19.75%
86	19.75%
17	17.28%
32	16.05%
6	9.88%
18	8.64%
45	8.64%
55	8.64%

A análise descritiva das classificações foi a seguinte:

Média	14.859
Máximo	17.200
Mínimo	11.000
P75	15.8
P50	15
P25	14.2
Reprovações	0

Na plataforma electrónica foram recepcionadas 311 contestações provenientes de 39 candidatos (45% do total) com a seguinte distribuição geográfica: Algarve (2); Centro (8); LVT (13); Madeira (1); Norte (15).
A Comissão Técnica responsável pela elaboração da prova reuniu, leu, apreciou e respondeu a todas as contestações nos dias 25 e 26 de Outubro, tendo em consequência deliberado anular as perguntas 4, 6, 25 e 82, aceitar 2 hipóteses corretas nas perguntas 24, 64 e 77 e alterar a chave da pergunta 55.

A alteração da chave foi enviada aos júris dos exames juntamente com a reclassificação dos candidatos a este teste e foram geradas novas pautas, agora definitivas das classificações das provas teóricas.

A média das classificações finais (após alteração da chave) aumentou. A análise descritiva das classificações finais é a seguinte:

Média	15.374
Máximo	18.125
Mínimo	11.458
P75	16.250
P50	15.625
P25	14.792
Reprovações	0

As 311 respostas da Comissão Técnica às contestações foram enviadas às Coordenações que posteriormente as endereçaram aos Candidatos (2 de Novembro).

Ocorrências

Após publicação pelos júris, nos dias 30 e 31 de Outubro, das classificações definitivas começaram a surgir exposições, pedidos de reunião e contestações às respostas e decisões da Comissão Técnica dirigidas a diferentes entidades (OM, Colégio de MGF, Coordenações). Algumas delas foram remetidas a esta comissão. De uma forma geral pretende-se contestar esta prova mediante o seguinte conjunto de afirmações a que iremos de imediato responder:

1) As perguntas com uma taxa de resposta certa <20% devem ser anuladas
2) As perguntas anuladas com taxa de respostas certas superiores a 98% prejudicam os candidatos devendo por isso dar-se um ponto a todos
3) Este teste é mais difícil do que o anterior, por isso a classificação deste teste deve ser ponderada.
4) As classificações do teste não podem ser tão diferentes das classificações das provas teóricas e curriculares
5) O teste deve ter uma forma diferente de cotação
6) O teste deve cumprir várias condições para ser um bom teste

Contextualização do problema

Da análise do conteúdo da maioria das contestações/exposições, sobressai a relação do impacto dos resultados do teste na seriação dos candidatos quando existir o concurso de colocação de especialistas em medicina geral e familiar, pois esse concurso será aberto para especialistas que foram avaliados em diferentes alturas, com testes de escolha múltipla diferentes e tendo efectuado outro tipo de provas.

Assim, importa relembrar o propósito do teste. O principal propósito do teste de escolha múltipla (TEM) é distinguir candidatos que têm os conhecimentos mínimos para serem reconhecidos como especialistas em medicina geral e familiar, dos que não têm. Dito de outra forma, o principal propósito é de certificação de candidatos e não de seriação de candidatos. É o entender da comissão técnica que este tipo de prova tem servido o objetivo de reconhecer quem se pode certificar como especialista em medicina geral e familiar.

Este teste de escolha múltipla veio substituir a anterior prova de conhecimentos, feita sob a forma de exame oral. Até ao princípio deste milénio o número de candidatos à obtenção do título de especialista era de 100 e neste momento o seu número anual é de cerca de 450. Com o TEM, os candidatos passaram a ter que demonstrar conhecimentos num leque muito mais alargado de conhecimentos, necessários à prática da medicina geral e familiar e todos os candidatos, na mesma época de saída, passaram a ser avaliados exactamente sobre os mesmos conteúdos, da mesma forma e nas mesmas condições. Por fim, o teste de escolha múltipla permitiu muito maior transparência no processo de avaliação.

O teste de escolha múltipla não foi desenhado para comparar candidatos avaliados em épocas diferentes. Existem variações na classificação média (e mediana) entre

candidatos avaliados em diferentes épocas. Utilizando a teoria clássica de teste, é Impossível distinguir se estas variações se devem a variações da dificuldade do teste ou a variações nos conhecimentos médios dos candidatos (1).

As perguntas com uma taxa de resposta certa <20% devem ser eliminadas?

As perguntas nas quais menos de 20% dos candidatos respondem correctamente são consideradas perguntas difíceis e as perguntas às quais mais de 80 % dos candidatos respondem podem ser consideradas de muito fáceis. O índice de dificuldade foi identificado como um dos métodos para determinar o nível de dificuldade de um exame, classificando-o em 3 níveis: fácil, moderado e difícil. O propósito do índice de dificuldade, é:

1. Identificar o conceito que precisa ser aprendido, pois descobre que os candidatos não conseguem responder a questões particulares (difícil).
2. Identificar e relatar a força e fraqueza das partes curriculares, que podem ou não podem ser dominadas pelos internos.
3. Dar informação de retorno aos gestores dos programas sobre as forças e fraquezas nos tópicos avaliados.
4. Identificar questões que são reveladoras de viés de conteúdo

Perguntas demasiado fáceis ou perguntas demasiado difíceis habitualmente têm pouca capacidade de discriminação dos candidatos. A análise da dificuldade das perguntas pode sinalizar perguntas mal formuladas, ou áreas que os candidatos não dominam adequadamente. No entanto, isto não significa que não devam existir perguntas muito fáceis ou difíceis. O National Medical Board of Medical Examiners (NBME) em 2016 refere que um teste de elevada qualidade inclui perguntas que cubram adequadamente os conteúdos que necessitam ser avaliados, que sejam representativas dos conhecimentos esperados num especialista de medicina geral e familiar e que representem um leque adequado de dificuldade. Dito de outra forma, não encontramos justificação teórica para eliminar as perguntas com maior grau de dificuldade.

A existir um problema com o nível de dificuldade das perguntas no TEM, esse é a existência de um número elevado de perguntas fáceis (habitualmente, 20% das perguntas da TEM têm um nível de dificuldade baixo). Este resultado não é surpreendente, uma vez que é lógico que a larga maioria dos candidatos a especialistas em medicina geral e familiar dominem as tarefas de diagnóstico, tratamento e gestão de doentes com os problemas habitualmente avaliados em consulta de medicina geral e familiar. As perguntas com menos de 20% ou mais de 80%, carecem ser revistas. Porém pode acontecer o problema não estar na pergunta, na sua formulação ou no conteúdo; mas, nos conhecimentos dos candidatos ou no programa.

As perguntas anuladas com taxa de respostas certas superiores a 98% prejudicam os candidatos

Alguns candidatos consideram que se deve dar um ponto quando se anula a pergunta, cotar a pergunta como se tivessem acertado porque de outro modo isso os prejudica.

Em geral, existe uma argumentação falaciosa, quando se afirma que se acertou a uma pergunta errada. Exemplifica-se da seguinte forma:

Pergunta: Diga qual dos seguintes valores é o resultado da operação 4x4 ?
a)17
b)18
c)19
d) 20

Na resposta a esta pergunta 96% dos respondentes assinalam a hipótese a). Só um respondente contesta a pergunta devido ao erro das hipóteses. Admitindo o erro a pergunta é anulada.

Ora a maioria dos respondentes não acertou a pergunta. A pergunta está errada, não deveria ter sido respondida. Não se pode acertar a pergunta errada a menos que esta afinal não esteja errada porque alguém ou a totalidade dos 96% dos respondentes defendam como certa a alínea a).
O conceito de prejuízo e injustiça não é aplicável pois esta situação não implica perda. Há uma redistribuição do valor da pergunta pelas demais.
O entendimento de prejuízo e de injustiça que esta argumentação dá a entender deve ser alvo de reflexão ética porque pressupõe que o benefício de um mal, de algo que está errado, se sobrepõe à correcção do erro ou à sua minimização.

Os resultados do TEM de Outubro de 2018 foram claramente inferiores aos dos TEM de Abril e deve-se proceder a uma ponderação

Este ponto é particularmente importante para esta discussão. Desde o início dos testes de conhecimentos através de perguntas de escolha múltipla, é possível verificar que há variações da classificação média ao longo dos exames de saída. Nalgumas épocas as variações são positivas (no sentido que um grupo de candidatos teve, em média, melhor classificação), noutras épocas as variações são negativas (no sentido que um grupo de candidatos teve, em média, pior classificação). Parte desta variação é explicada por variação na dificuldade dos testes, parte desta variação é explicada por variações nos conhecimentos médios dos candidatos.

A comissão técnica admite que uma parte desta variação se deva a variações da dificuldade do teste, mas desconhece técnicas que permitam isolar quanto desta variação é devida à dificuldade do teste e quanto desta variação é devida a diferenças entre candidatos.

No TEM de Abril de 2018 existiram 2 perguntas em que menos de 30% dos candidatos responderam correctamente enquanto no teste de Outubro existiram 5 (após revisão).

As classificações do teste não podem ser tão diferentes das classificações das provas teóricas e curriculares

Em defesa desta tese os candidatos argumentam que "*as notas das Provas de discussão curricular e das Provas práticas, anteriormente realizadas à Prova teórica, não diferem significativamente das épocas anteriores. Ora, se* a *discussão curricular e os conhecimentos necessários para realização de uma prova prática se constroem sobre conhecimentos teóricos, não serão coerentes os resultados deste TEM*".

Antes de existir TEM, as notas das Provas de discussão curricular, das Provas práticas e das Provas teóricas não diferiam significativamente entre si, criando a sensação que as notas eram niveladas homogeneamente, não existindo nenhuma prova comum que permitisse a distinção entre os candidatos. Esta prova, que agora alguns candidatos discutem, vem colmatar essa lacuna na avaliação, colocando pelo menos uma prova igual para todos.
As Provas de discussão curricular e as Provas práticas são provas qualitativas cuja apreciação é erradamente traduzida para uma escala quantitativa de valores referência e não para uma escala de 0 a 20 dada isso não ser possível. Só se pode comparar o que é comparável.

O teste deve ter uma forma diferente de cotação

Alguns candidatos defendem que "*As questões clinicamente mais aplicáveis, relativas à selecção de uma "melhor opção", apresentam um cenário com* ***o****pções relevantes e plausíveis …; a resposta "melhor" pode ser julgada como 80% correta e os distratores talvez 20 a 30% correctos…*".(2)

A forma de cotação das perguntas pode ser discutível. Sobre as cotações dos testes há diferentes opções. Foi decidido que este teste teria 100 perguntas de 4 opções, com uma única resposta certa. Esta opção poderá ser revista no futuro e a Comissão que elabora a prova terá isso em consideração. Num dos artigos citados pelos signatários de uma das exposições McCoubrie, P. (2004) é claramente referido " There is no single recommended approach to setting standards".

Tem havido um esforço muito grande da comissão técnica em criar perguntas unidimensionais, com uma opção correta e distratores plausíveis. Tem havido também um grande esforço para privilegiar perguntas com conteúdo relevante para a prática dos especialistas em medicina geral e familiar, em detrimento de perguntas que possam ser respondidas apenas memorizando o conteúdo de orientações clínicas. Curiosamente, são as perguntas clínicas, unidimensionais e clinicamente relevantes que são as mais frequentemente contestadas.

O teste deve cumprir várias condições para ser um bom teste

Os candidatos socorrem-se de 2 textos (1,2) sobre testes de perguntas de escolha múltipla para defender que o teste deve " *Ser alinhado e coerente com os objectivos de aprendizagem; II. Ser combinado com testes de competências práticas; III. Traduzir uma amostra representativa do conteúdo importante; IV. Estar isento de itens contendo erros e falhas técnicas; V. Utilizar um critério de aprovação/retenção baseado em critérios (absoluto) e não em normas (relativo)***.** *VI. Ser cuidadoso na reutilização de perguntas de teste; VII. Utilizar formatos alternativos que já provaram a sua validade e fidedignidade (p.ex., Extended-Matching)*".(3)

Há imensa literatura sobre perguntas de escolha múltipla. Uma breve pesquisa no motor de pesquisa generalista Google Scholar apresenta 1 460 000 resultados (em 0,12 segundos). Talvez a mais divulgada seja a publicada pelo National Board of Medical Examiners ou a publicada pela AMEE (European Association of Medical Education) que a Comissão que elabora a prova conhece (3,4). A maioria dos elementos da Comissão Técnica tem razoável experiência na elaboração de testes de perguntas de escolha múltipla e formação específica nesta área e julga que os Testes, no geral, têm cumprido as condições acima. Os diversos Testes, mais do que qualquer outra prova de avaliação, têm sido submetidos a um enorme escrutínio, são muito mais abrangentes nas matérias que avaliam, conseguem ter uma forte ligação à prática clínica e avaliar várias componentes do conhecimento: o conhecimento de factos, interpretação, decisão, aplicação de conhecimentos. Existirem perguntas que após contestações são anuladas ou às quais se muda a chave não desqualifica o teste embora seja desejável que tal não aconteça.

Conclusão

O Teste de Escolha Múltipla justifica-se pelo propósito a que se destina não lhe devendo ser dado propósito diferente.

A distribuição das classificações de um teste resulta da qualidade das perguntas do teste e dos atributos dos candidatos.

As perguntas difíceis assim como as perguntas muito fáceis não são de eliminar liminarmente. A Comissão Técnica entende que a eliminação de perguntas baseada no

grau de dificuldade levantará a dificuldade prática de atribuir classificações abaixo de 9,5 valores a candidatos que foram já aprovados.

A aplicação de fatores de correcção nas notas obtidas no concursos de saída, a acontecer, terá que ser aplicada a outros concursos. Isto significa que haverá especialistas em medicina geral e familiar que verão a sua nota "subir" para fim de provimento; haverá especialistas em medicina geral e familiar de outros concursos que verão a sua nota descer. Se os fatores de correcção não forem aplicados a todos, haverá especialistas que serão duplamente prejudicados (isto é, pertencerão a concursos que em média tiveram classificações mais baixas e não pertencerão a um concurso em que houve 'correcção' das notas).

É desejável que este tipo de teste de certificação inclua cada vez mais perguntas com enfoque em níveis mais elevados de cognição, conforme definido pela taxonomia de Bloom. Ou seja que a vinheta ou tronco da pergunta apresente um problema que requeira a aplicação de princípios, a análise de um problema ou a avaliação de alternativas ou que requeira pensamento multilógico. Este tipo de perguntas com alternativas que exigem um alto nível de discriminação irão contribuir para que este teste tenha mais contestações sendo por isso necessário pensar-se qual a melhor estratégia a adoptar ou qual o tipo de teste desejado.

É imprescindível manter o cuidado tido na elaboração da prova sendo que a manter-se o número de candidatos e o volume de contestações é necessário profissionalizar, alargar e apoiar a Comissão Técnica (CT). A CT precisa de apoio informático institucional e horas de trabalho dedicadas a estas tarefas.

O teste de escolha múltipla deve ser testado antes de ser aplicado.

A bibliografia recomendada de suporte à prova deve ser actualizada e revista.

Seria aconselhável que nos concursos de provimento entrassem outros componentes de avaliação dos candidatos que não exclusivamente a classificação, como aliás já acontece mas só para alguns (exemplo: escolhas dirigidas para ingresso em USF modelo B).

Não parece existir qualquer situação de injustiça proporcionada pelo teste pois todos os candidatos dispõem das mesmas condições para o efectuar. A injustiça sentida estará no modelo de provimento dos candidatos que ficam sujeitos a 2 modelos de

selecção: um exclusivamente pela classificação de saída em que ficam seriados e outro só por entrevista.

A categorização e a validação dos especialistas deveriam ser feitas por níveis de mérito e não por uma classificação numérica, pois não se dispõe de instrumentos de avaliação que na globalidade possam ter esta métrica. Não é adequado classificar de 0 a 20 uma prova curricular ou uma prova prática pois estas provas como já anteriormente referido são qualitativas. Não existem médicos de 18,6 ou de 18,7.

Nota: Sugere-se que este relatório seja divulgado pelas instituições competentes a: Comissão de Internos, APMGF, CNPG/OM

Comissão Técnica que elabora a prova: Bruno Heleno, Carla Correia, Catarina Matias, Fernando Ferreira, Isabel Santos, José Mendonça, Luís Alves, Luísa Sá, Maria da Luz Loureiro, Silva Henriques

Criação e gestão das plataformas electrónicas: David Rodrigues, Cecília Shinn

Lisboa, 8 de Novembro de 2018

Bibliografia

1 - Streiner DL Norman GR Cairney J Health Measurement Scales – a practical guide to their development and use. 5th Edition. Oxford: Oxford University Press; 2015

2 -NBME. Constructing Written Test Questions For the Basic and Clinical Sciences, 3 ed rev . National Board of Medical Examiners 3750 Market Street Philadelphia, PA 19104 https://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf

3- McCoubrie, P.. Improving the fairness of multiple-choice questions: a literature review. Medical Teacher, Vol. 26, No. 8, 2004, pp. 709–712

4 - Walsh JL, Harris BHL, Smith PE. Single best answer question-writing tips for clinicians. BMJ Postgrad Med J 2017;93:76–81.

5 - BANDARANAYAKE, RAJA C. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. MEDICAL TEACHER 2008; 30: 836–845

6 - Vanderbilt University. Centre for teaching https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/#guidelines