INOValiar: lista consolidada de critérios de qualidade de questões de escolha múltipla na avaliação de aprendizagens

JOÃO PADILLA[I]
MARTA MATEUS DE ALMEIDA[II]
CARLA NASCIMENTO[III]
MARA PEREIRA GUERREIRO[IV]
http://dx.doi.org/10.22347/2175-2753v17i54.5021

Resumo
As questões de escolha múltipla (QEM) são amplamente utilizadas na avaliação de estudantes do ensino superior, com capacidade de avaliar diversos domínios cognitivos. No entanto, a literatura científica indica que um número considerável de QEM apresenta não conformidades face às orientações de boa prática, comprometendo a validade e fiabilidade do processo de avaliação. Este trabalho foca-se na componente técnica da construção de QEM de resposta única, apresentando uma lista consolidada de 18 critérios de qualidade que derivam da literatura. Este recurso foi utilizado na formação de 176 docentes do ensino superior de várias disciplinas, que expressaram opiniões positivas e manifestaram a intenção de integrá-lo nas suas práticas de avaliação.
Palavras-chave: questões de escolha múltipla; avaliação de estudantes; critérios de boa prática; ensino superior.

Submetido em: 01/08/2024
Aprovado em: 20/03/2025

[I] Escola Superior de Enfermagem de Lisboa (ESEL), Lisboa, Portugal; https://orcid.org/0000-0002-5689-0909; e-mail: joao-padilla@esel.pt.
[II] Instituto de Educação da Universidade de Lisboa (UIDEF), Lisboa, Portugal; https://orcid.org/0000-0003-3108-4289; e-mail: mialmeida@ie.ulisboa.pt.
[III] Escola Superior de Enfermagem de Lisboa (ESEL), Lisboa, Portugal; https://orcid.org/0000-0002-4880-0141; e-mail: carla.nascimento@esel.pt.
[IV] Egas Moniz School of Health and Science, Almada, Portugal; https://orcid.org/0000-0001-8192-6080; e-mail: mguerreiro@egasmoniz.edu.pt.

INOValiar: consolidated list of quality criteria for multiple-choice questions in learning assessment

Abstract
Multiple-choice questions (MCQs) are widely used in the assessment of higher education students, with the ability to evaluate various cognitive domains. However, the scientific literature indicates that many MCQs do not conform to best practice guidelines, compromising the validity and reliability of the assessment process. This work focuses on the technical component of constructing one-best-answer MCQs, presenting a consolidated list of 18 quality criteria derived from the literature. This resource was used in the training of 176 higher education faculty members from multiple disciplines, who expressed positive opinions and indicated their intention to integrate it into their assessment practices.
Keywords: multiple-choice questions; student assessment; best practice criteria; higher education.

INOValiar: lista consolidada de criterios de calidad para preguntas de opción múltiple en la evaluación del aprendizaje

Resumen
Las preguntas de opción múltiple (POM) se utilizan ampliamente en la evaluación de estudiantes de educación superior, con la capacidad de evaluar diversos dominios cognitivos. Sin embargo, la literatura científica indica que un número considerable de POM presenta inconformidades con respecto a las directrices de buenas prácticas, comprometiendo la validez y fiabilidad del proceso de evaluación. Este trabajo se centra en el componente técnico de la construcción de POM de respuesta única, presentando una lista consolidada de 18 criterios de calidad derivados de la literatura. Este recurso se utilizó en la formación de 176 docentes de educación superior de varias disciplinas, quienes expresaron opiniones positivas y manifestaron su intención de integrarlo en sus prácticas de evaluación.
Palabras-clave: preguntas de opción múltiple; evaluación de estudiantes; criterios de buenas prácticas; educación superior.

Introdução

A avaliação ainda continua a ser um dos fatores mais significativos que influencia a aprendizagem dos estudantes no ensino superior (Bryan; Clegg, 2006). Ao longo dos anos, a avaliação tem desempenhado funções de dimensão social e de dimensão pedagógica (Santos, 2019).

As preocupações com a validade e utilidade das informações recolhidas nos processos avaliativos e, por conseguinte, na construção dos instrumentos de avaliação, tem sido um dos domínios abordados na literatura especializada (Fernandes, 2019; Fernandes; Borralho; Barreira; Monteiro; Catani; Cunha; Alves, 2014; Fernandes; Machado; Abelha, 2023; Santos; Pinto, 2018). A procura de maior objetividade e credibilidade dos instrumentos deriva numa perspetiva mais técnica de definição de regras e enquadra-se numa visão do paradigma psicométrico (Kellaghan; Madaus, 2000; Scriven, 2000), sem que com isto se reduza a avaliação a uma visão meramente tecnicista.

As questões de escolha múltipla (QEM) são um instrumento amplamente popularizado na avaliação de aprendizagens do domínio cognitivo (Bollela; Borges; Troncon, 2018; Bredon, 2003), que surgiu na década de 1910, nos Estados Unidos da América, como resposta à massificação do ensino e ao facto de se detetarem grandes discrepâncias na correção de exames com perguntas do tipo ensaio (Fernandes, 2004).

Como instrumento de avaliação, as QEM destacam-se pela objetividade, imparcialidade e possibilidade de automatização de cotação e *feedback* (Brown; Abdulnabi, 2017; Coughlin; Featherstone, 2017; Dell; Wantuch, 2017; Gupta; Williams; Wadhwa, 2021). Podem ser usadas no contexto da avaliação sumativa (dimensão social), permitindo a classificação do desempenho dos estudantes no final do período de aprendizagem, ou na avaliação formativa (dimensão pedagógica), facultando aos estudantes informação sobre a sua aprendizagem. Inclui-se nesta última dimensão a utilização de QEM no âmbito de práticas fundamentadas pela evidência, como a prática da recuperação (Hattie; Donoghue, 2016). Independentemente do âmbito da sua utilização, se redigidas adequadamente, as QEM têm a capacidade de avaliar vários domínios cognitivos, inclusivamente de natureza superior, como aplicação e análise, segundo a taxonomia revista de Bloom (Dell; Wantuch, 2017; Gupta; Williams; Wadhwa, 2021; Smith, 2018).

No entanto, vários trabalhos têm apontado não conformidades face às orientações de boa prática nas QEM usadas na avaliação de estudantes (Downing, 2005; Fayyaz Khan; Farooq Danish; Saeed Awan; Anwar, 2013; Nedeau-Cayo; Laughlin; Rus; Hall, 2013; Pais; Silva; Guimarães; Povo; Coelho; Silva-Pereira; Lourinho; Ferreira; Severo, 2016; Tarrant; Ware, 2008). Por exemplo, Tarrant e Ware (2008) analisaram 664 QEM usadas na avaliação sumativa de estudantes de enfermagem e concluíram que 47,3% possuíam não conformidades face às recomendações para redação de QEM. Um estudo português (Pais; Silva; Guimarães; Povo; Coelho; Silva-Pereira; Lourinho; Ferreira; Severo, 2016) identificou o mesmo problema em cerca de metade das 800 QEM (55,8%) usadas na avaliação de estudantes de medicina.

Considerando que a validade e a fiabilidade são duas das principais características psicométricas do processo de avaliação (Kellaghan; Madaus, 2000), desenvolver QEM adequadas e baseadas nas orientações de boa prática é crucial para garantir a qualidade e a fiabilidade do processo de avaliação (Catanzano; Jordan; Lewis, 2022; Dell; Wantuch, 2017; Tarrant; Ware, 2008). Uma QEM inadequadamente redigida pode aumentar a dificuldade de compreensão do enunciado ou fornecer pistas que tornam a resposta correta mais evidente (National Board of Medical Examiners, 2021). Por exemplo, o trabalho de Pais, Silva, Guimarães, Povo, Coelho, silva-Pereira, Lourinho, Ferreira e Severo (2016) indica que não conformidades no enunciado e nas opções de resposta contribuem para maior dificuldade e menor discriminação das questões.

O índice de dificuldade baseia-se na proporção de estudantes que respondem corretamente à questão face ao número total de estudantes que responderam à questão (Gupta; Williams; Wadhwa, 2021; Morgado *et al*., 1997; National Board of Medical Examiners, 2021; Tavakol; Dennick, 2011). Tipicamente, considera-se adequado um intervalo de dificuldade entre 30 e 80, sendo que QEM com valores abaixo de 30 e acima de 80 são considerados difíceis e fáceis, respetivamente (Morgado *et al*., 1997; Salih; Jibo; Ishaq; Khan; Mohammed; Al-Shahrani; Abbas, 2020; Tavakol; Dennick, 2011).

O índice de discriminação reflete a capacidade de uma questão diferenciar os estudantes, considerando a classificação obtida na prova (Gupta; Williams; Wadhwa, 2021; Morgado *et al*., 1997; National Board of Medical Examiners, 2021). Apesar da literatura educacional não ser unânime, existe uma convergência nos pontos de corte deste indicador psicomético. O trabalho de Salih, Jibo, Ishaq, Khan,

Mohammed, Al-Shahrani e Abbas (2020), por exemplo, indica que valores acima de 0,40 refletem QEM muito boas, entre 0,30 e 0,39 razoavelmente boas, entre 0,20 e 0,29 QEM marginais, que devem ser sujeitas a melhoria, e abaixo de 0,19 QEM que devem ser rejeitados ou revistos.

A Agência de Avaliação e Acreditação do Ensino Superior (A3ES) é a entidade responsável pelo processo de avaliação e acreditação das instituições de ensino superior e dos seus ciclos de estudos em Portugal. Com o objetivo de garantir a qualidade do ensino superior, a A3ES definiu um conjunto de referenciais, que orientam as instituições no processo de implementação e acreditação dos sistemas internos de garantia da qualidade (SIGQ). O Referencial 3 preconiza que as instituições de ensino superior estabeleçam mecanismos para garantir que o processo de avaliação de estudantes é adequado, consistente e alinhado com os objetivos de aprendizagem previamente definidos (Agência de Avaliação e Acreditação do Ensino Superior, 2016).

Assim, torna-se imprescindível compreender as questões técnicas inerentes ao desenvolvimento dos instrumentos de avaliação de estudantes e promover boas práticas na elaboração de QEM, com vista a assegurar a qualidade e a fiabilidade do processo de avaliação. Sem se pretender reduzir a avaliação a uma visão meramente tecnicista e cientes dos problemas amplamente discutidos na literatura (Fernandes; Fialho, 2012), este trabalho centra-se na componente técnica da construção de instrumentos de avaliação constituídos por QEM e pretende contribuir para a garantia de qualidade na avaliação de aprendizagens.

Embora existam vários trabalhos com recomendações para a construção de QEM, não identificámos recursos em língua portuguesa, assentes num processo sistemático de desenvolvimento com base na literatura ténico-científica. Este trabalho visa dar resposta a esta necessidade, tendo como objetivo desenvolver de uma lista consolidada de critérios para garantir a qualidade na avaliação de aprendizagens através de QEM de resposta única.

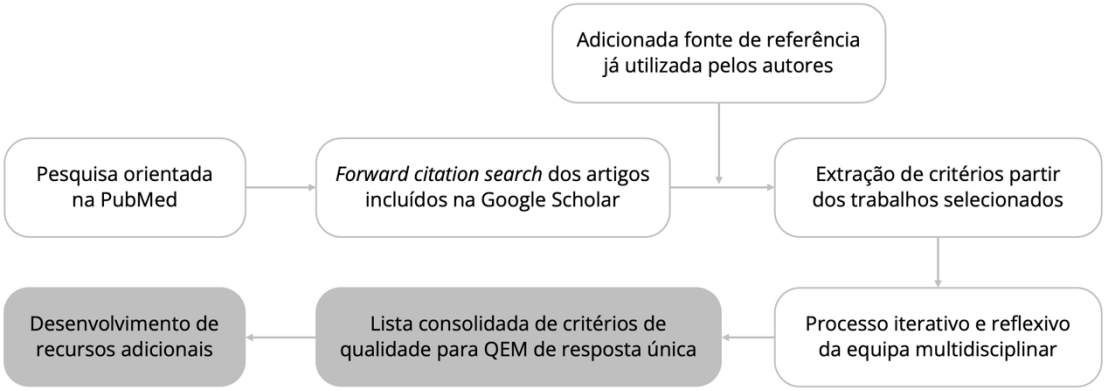
Metodologia

Pesquisa bibliográfica

Foi realizada uma pesquisa orientada na base de dados *PubMed* para identificar trabalhos que abordassem práticas recomendadas na elaboração de QEM. Conforme descrito na Figura 1, a estratégia de pesquisa focou-se em artigos

publicados entre 2017 e 2023, cujo título incluía a expressão *multiple choice*. Não foram aplicadas restrições adicionais à pesquisa. Para garantir a abrangência dos resultados, foi realizada uma pesquisa complementar de citações das referências selecionadas via *Google Scholar*. Foram incluídos todos os trabalhos em língua inglesa ou portuguesa que discutissem orientações explícitas para a construção de QEM. Adicionalmente, integrou-se uma fonte de referência previamente utilizada pelos autores (National Board of Medical Examiners, 2021).

Figura 1 - Fluxograma da metodologia implementada



Fonte: Os autores (2024).

Extração e consolidação de dados

A partir dos trabalhos selecionados, extraiu-se uma lista preliminar de critérios de qualidade. Essa lista foi submetida a um processo iterativo de análise e consolidação por uma equipa multidisciplinar, visando congregar as orientações dispersas pelos vários trabalhos numa lista consolidada.

Desenvolvimento de recursos adicionais

Antecipando o uso da lista consolidada, e com base na experiência da equipa em educação, investigação e inovação, desenvolveram-se recursos adicionais para apoiar as questões técnicas inerentes ao desenvolvimento de instrumentos de avaliação baseados em QEM de resposta única. Entre estes, um glossário para harmonizar a terminologia específica e facilitar a utilização da lista consolidada.

Formação e *feedback* de docentes do ensino superior

Foram realizadas cinco sessões de formação sobre instrumentos de avaliação com QEM de resposta única para docentes do ensino superior. Quatro dessas sessões, com duas horas de duração cada, foram realizadas à distância. A outra sessão adotou uma abordagem híbrida, com três horas de formação presencial e uma hora de acompanhamento à distância, com uma semana de intervalo. No total, participaram 176 docentes de diversas disciplinas, que aplicaram a lista consolidada em várias QEM de resposta única com não conformidades face aos critérios de qualidade.

No final das sessões, foi recolhida a opinião dos docentes através de questões de resposta fechada e aberta, administradas por meio de um sistema de votação eletrónica, com o intuito de aferir a usabilidade das ferramentas desenvolvidas e antever a sua adequação às necessidades práticas dos docentes.

Resultados

Pesquisa bibliográfica

Foram identificados 218 trabalhos na base de dados *PubMed*. A leitura dos títulos e resumos resultou na seleção de nove trabalhos. Contudo, um dos trabalhos foi excluído devido à falta de acesso ao texto integral. A pesquisa complementar de citações posteriores dos artigos selecionados resultou na inclusão de dois trabalhos adicionais. Além disso, integrou-se uma fonte de referência previamente utilizada pelos autores (National Board of Medical Examiners, 2021). No total, foram incluídos onze trabalhos, a partir dos quais foi possível extrair mais de quarenta critérios de qualidade.

Critérios de qualidade

Após uma análise detalhada e comparação iterativa pela equipa multidisciplinar, consolidaram-se 18 critérios de boa prática para a elaboração de QEM de resposta única. Todos os critérios foram redigidos sobre a forma de pergunta, cuja resposta desejável é "sim", a fim de facilitar a implementação (Tabela 1).

Os critérios foram estratificados em três categorias, refletindo diferentes elementos técnicos das QEM de resposta única:

- Grupo A (2 critérios): foco da questão
- Grupo B (4 critérios): enunciado

- Grupo C (11 critérios): opções de resposta

No Grupo A, que incide sobre o foco da questão, os critérios enfatizam a importância do alinhamento construtivo e do número de resultados de aprendizagem específicos avaliados por cada questão. De acordo com os princípios do alinhamento construtivo (Biggs, 1996), os instrumentos de avaliação devem estar estritamente alinhados com os resultados de aprendizagem previamente definidos, a fim de garantir a qualidade do processo pedagógico. Isto permite avaliar precisamente os conhecimentos ou habilidades que os estudantes devem adquirir, sendo esse o escopo do critério A1. Por sua vez, o critério A2 visa assegurar que cada QEM se concentra em apenas um resultado de aprendizagem específico, evitando a diluição do foco e melhorando a precisão na avaliação das competências dos estudantes.

Tabela 1 - Lista consolidada com critérios de qualidade para elaboração de QEM de resposta única

	Critério	**Resposta desejável**
	Grupo A: foco da questão de escolha múltipla	
A1	A QEM está alinhada com os objetivos/resultados de aprendizagem previamente estabelecidos?	Sim
A2	A QEM foca-se em apenas um objetivos/resultados de aprendizagem específico?	Sim
	Grupo B: enunciado	
B1	A redação do enunciado é compreensível aquando da primeira leitura (ex. evita palavras de difícil entendimento)?	Sim
B2	O enunciado é redigido sem recurso a advérbios de negação ou exclusão? (ex. "não", "menos" ou exceto")	Sim
B3	Termos de frequência vagos são omissos no enunciado? (ex. "geralmente", "principalmente", "frequentemente" ou "raramente")	Sim
B4	O enunciado fornece apenas a informação necessária para responder à pergunta?	Sim
	Grupo C: opções de resposta	
C1	A redação das opções de resposta é compreensível e concisa? (ex. não é necessário ler mais do que uma vez a opção de resposta para perceber o que está escrito; não existem palavras acessórias para a compreensão)	Sim
C2	As opções de resposta são apresentadas de forma ordenada e coerente? (ex. ordenar de forma lógica as opções de resposta ou todas as opções numéricas são apresentadas em % ou em decimal)	Sim

Continua

Conclusão

Critério		Resposta desejável
Grupo C: opções de resposta		
C3	As opções de resposta têm, aproximadamente, a mesma extensão? (ex. as opções de resposta são aproximadamente semelhantes em número de palavras)	Sim
C4	As opções de resposta são homogéneas em termos de conteúdo (ex. avaliam a mesmo assunto, e não outros aspetos distintos)	Sim
C5	As opções de resposta têm uma estrutura gramatical coerente com o enunciado (ex. são evitadas pistas gramaticais)?	Sim
C6	Estão ausentes as expressões "nenhuma das anteriores" ou "todas as anteriores" nas opções de resposta?	Sim
C7	Os distratores utilizados são plausíveis? (ex. nenhum distrator deve ser excluído sem requerer análise por parte do respondente)	Sim
C8	Termos de frequência vagos são omissos nas opções de resposta? (ex. "geralmente", "principalmente", "frequentemente" ou "raramente")	Sim
C9	As opções de resposta são isentas de termos absolutos? (ex. "sempre, "nunca" ou "todos")	Sim
C10	As opções de resposta são independentes e não se sobrepõem? (ex. uma opção de resposta não deve conter informação que exista noutras opções)	Sim
C11	As várias opções de resposta são isentas de repetição de termos que permita convergir para a resposta certa?	Sim
C12	É evitada a utilização dos mesmos termos ou de sinónimos na opção de resposta correta e no enunciado?	Sim

Fonte: Os autores (2024).

No Grupo B estão incluídos critérios que visam evitar a introdução de dificuldade irrelevante no processo de avaliação dos estudantes. Um exemplo disso é o critério B2, que determina que o enunciado deve ser formulado sem o uso de advérbios de negação ou exclusão, como "não", "menos" ou "exceto". A formulação negativa aumenta desnecessariamente a complexidade cognitiva dos estudantes. Exemplos de perguntas formuladas negativamente podem ser consultados no Quadro 1.

Quadro 1 - Exemplos de perguntas formuladas de forma negativa

Qual destes autores NÃO é considerado um representante do Romantismo?
Todos os seguintes são princípios da dinâmica, EXCETO:
Qual das seguintes NÃO é uma característica fundamental do capitalismo?
Qual das seguintes equações NÃO representa uma linha reta?
Todos os seguintes são exemplos de reações exotérmicas, EXCETO:
Qual das seguintes opções NÃO é uma função da membrana celular?

Fonte: Os autores (2024).

O Grupo C inclui uma variedade de critérios focados nas opções de resposta, tais como a coerência gramatical, a homogeneidade em termos de conteúdo e extensão e a organização lógica das opções numéricas. Por exemplo, o critério C2 sublinha que as opções de resposta numéricas devem ser apresentadas de forma ordenada e coerente para evitar dificuldade irrelevante ao estudante. Isto requer que todas as opções de resposta tenham a mesma unidade (ex. centímetros ou percentagem) e organizadas de forma lógica. O critério C3 estipula que as diferentes opções de resposta devem ter aproximadamente a mesma extensão, evitando dar pistas sobre a resposta correta através do número de palavras. Habitualmente, verifica-se esta não conformidade quando o docente se foca na opção de resposta correta, podendo incluir informação adicional que auxilia o estudante a identificar a resposta certa (National Board of Medical Examiners, 2021). Já o critério C5 requer que as opções de resposta mantenham uma estrutura gramatical coerente com o enunciado, para evitar pistas sobre opções de respostas erradas devido a inconsistências gramaticais. Por exemplo, uma pergunta que termina com uma expressão no feminino e possui opções de resposta no masculino pode permitir que o estudante exclua imediatamente essas opções de resposta (Quadro 2).

Quadro 2 - Exemplo de questão de escolha múltipla com não conformidade na coerência gramatical entre a pergunta e as opções de resposta

A indapamida é um fármaco cujo mecanismo de ação consiste **na**:
A. Redução da reabsorção de sódio e água no nefrónio.
B. Antagonismo dos recetores D2 no aparelho justaglomerular.
C. Ativação dos adrenorreceptores beta-1cardiosseletivos.
D. Bloqueio do influxo de cálcio no músculo liso vascular.

Fonte: Os autores (2024).

O critério C6 destaca uma não conformidade relativamente comum, e sublinha que expressões como "nenhuma das anteriores" ou "todas as anteriores" devem ser evitadas nas opções de resposta, pois podem complicar ou simplificar excessivamente a decisão do estudante. Por exemplo, quando a resposta mais correta não é "nenhuma das anteriores", os estudantes mais informados são colocados num dilema, pois têm de decidir entre a opção que o docente pretendeu como correta e uma opção que abrange tudo o que não está listado no conjunto de opções de resposta (National Board of Medical Examiners, 2021).

O critério C10 enfatiza que as opções de resposta devem ser independentes e não se sobreporem, para evitar introduzir dificuldades irrelevantes no instrumento de avaliação. Por exemplo, numa pergunta numérica em que as opções de resposta são:

A. 1-5
B. 5-10
C. 10-15
D. 15-20

existe sobreposição em algumas opções de resposta e pode dificultar a decisão do estudante que acredita que a resposta exata é "5", "10" ou "15" (Catanzano; Jordan; Lewis, 2022).

O critério C11 enfatiza a importância de que as opções de resposta sejam formuladas de modo a evitar repetições de termos, que podem inadvertidamente facilitar a escolha da resposta correta para o estudante. Isso ocorre porque a repetição pode servir de pista, especialmente quando o estudante reconhece um padrão nas opções apresentadas. Um exemplo ilustrativo de uma não conformidade deste critério é a seguinte questão: quais dos seguintes antidiabéticos deve ser administrado por via subcutânea?

A. Insulina e metformina.
B. Dapagliflozina e dulaglutido.
C. Pioglitazona e insulina.
D. Insulina e dulaglutido. [CORRETA]

A presença recorrente da palavra "insulina" em três das opções de resposta pode levar os estudantes a inferir incorretamente que qualquer opção que inclua "insulina" é mais provável de ser a correta, baseando-se mais na frequência de aparição do termo do que no entendimento correto das indicações dos medicamentos. Para corrigir essa não conformidade, se for necessário que as opções de resposta sejam apresentadas em pares, é necessário garantir que cada termo apareça um número igual de vezes ao longo das opções:

A. Insulina e metformina.
B. Insulina e dulaglutido. [CORRETA]
C. Dulaglutido e pioglitazona.
D. Metformina e pioglitazona.

Recursos adicionais

Para facilitar a utilização da lista consolidada e apoiar a aplicação prática dos critérios de qualidade, desenvolveu-se um conjunto de recursos adicionais. Um desses recursos é um glossário (Tabela 2) que define a terminologia associada aos componentes das QEM de resposta única.

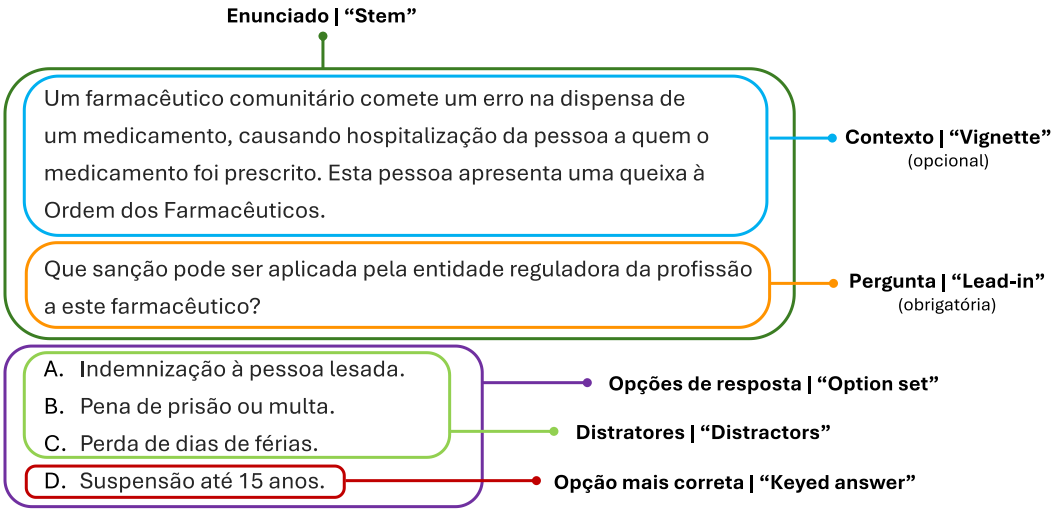
Tabela 2 - Glossário de terminologia sobre QEM

Termo	Definição
QEM de resposta única	Item de avaliação constituído por um enunciado e por um conjunto de opções de resposta, em que apenas uma é a mais correta.
Enunciado	Elemento constitutivo de uma QEM, que compreende um contexto (ex. cenário ou caso clínico), opcional, e obrigatoriamente uma pergunta.
Pergunta	Elemento constitutivo de uma QEM, que compreende uma pergunta, idealmente na forma interrogativa.
Distratores	Elementos constitutivos de uma QEM, que constituem respostas menos corretas ao enunciado.
Opção correta	Elemento constitutivo de uma QEM, que constitui a resposta mais correta ao enunciado.

Fonte: Os autores (2024).

Adicionalmente, foi criada um apoio visual que ilustra os elementos constituintes de uma QEM do tipo A (Figura 2).

Figura 2 - Exemplo da aplicação da terminologia sobre QEM



Fonte: Os autores (2024).

Feedback das sessões de formação

Globalmente, os participantes das sessões de formação expressaram opiniões muito positivas sobre a utilidade das ferramentas desenvolvidas para as suas práticas de avaliação. Por exemplo, numa das instituições, sete dos 10 participantes manifestaram intenção de reformular as suas atividades a curto prazo para implementar o que aprenderam. Numa outra formação, a totalidade dos 27 respondentes afirmou que gostaria de usar a lista consolidada na sua prática docente, sugerindo aceitabilidade desta ferramenta.

Estes resultados foram complementados com comentários que corroboram a perceção de utilidade e perspetivam o potencial destas ferramentas impactarem positivamente as práticas pedagógica, tais como: "a utilização desta ferramenta irá permitir detetar/diminuir as não conformidades nas QEM" e "excelente para refletir sobre a minha/nossa prática QEM".

Em termos operacionais, os docentes que participaram nas sessões de formação apontaram sugestões de melhoria para implementar na lista consolidada (Quadro 3), designadamente a inclusão de exemplos práticos que acompanhem cada critério para facilitar a interpretação a transposição para o contexto individual. Estes exemplos podem ser pertinentes para interpretar os critérios C10 e C11 que, reiteradamente, foram apontados pelos docentes como sobrepostos ou de difícil compreensão.

Quadro 3 - Exemplos de sugestões de melhoria apontadas por docentes do ensino superior

Penso que é sempre útil haver exemplos a seguir a cada questão, para que, numa primeira utilização, se consiga confirmar que a nossa interpretação está alinhada com o exemplo. Estes exemplos podem também ficar ocultos (acessível com um clique num ícone definido de "*help*" ou "?", por exemplo).
Era importante ter mais exemplos na lista, para que se consiga perceber melhor os conceitos.
Exploração de exemplos que foram enviados previamente ao formador.

Fonte: Os autores (2024).

Validade psicométrica

Considera-se que a lista consolidada de critérios de qualidade para elaboração de QEM de resposta única apresenta validade de face e conteúdo, atendendo à sua génese na literatura técnico-científica. Adicionalmente, os testemunhos dos docentes que participaram nas sessões de formação corroboram

essa inferência. Por exemplo, na primeira formação, 90% dos docentes afirmaram não ver necessidade de eliminar ou modificar qualquer critério, existindo dados textuais neste sentido: "a lista pareceu-me bastante interessante e todos os itens fazem sentido".

Perspetiva-se prosseguir à avaliação psicométrica desta ferramenta, através da aferição da validade de critério e de construto. Adicionalmente, pretende-se realizar a análise dos componentes principais (Grupo A, B e C), e determinar a fiabilidade da lista consolidada.

Conclusões

Alicerçado na literatura científica, desenvolveu-se uma lista consolidada com critérios de qualidade para elaboração de QEM de resposta única que se espera que contribuam para que o processo de avaliação de estudantes seja adequado, consistente e alinhado com os objetivos de aprendizagem previamente definidos, tal como preconizado pela A3ES (Agência de Avaliação e Acreditação do Ensino Superior, 2016).

A utilização desta lista consolidada em sessões de formação de docentes do ensino superior indicou elevada aceitabilidade e utilidade, levando também à identificação de oportunidades de melhoria. Entre estas oportunidades de melhoria, encontra-se a necessidade de incluir exemplos concretos que auxiliem na interpretação de cada critério, facilitando a aplicação prática.

Globalmente, esta ferramenta foi desenvolvida com o intuito de enriquecer a prática pedagógica e promover uma melhoria contínua dos instrumentos de avaliação baseados em QEM de resposta única. Espera-se que encorajem os docentes do ensino superior a refletir e aperfeiçoar constantemente as suas práticas pedagógicas.

Referências bibliográficas

AGÊNCIA DE AVALIAÇÃO E ACREDITAÇÃO DO ENSINO SUPERIOR. *Referenciais para os sistemas internos de garantia da qualidade nas instituições de ensino superior*: adaptado aos ESG 2015: versão 1.2. Lisboa: A3ES, 2016.

BIGGS, J. Enhancing teaching through constructive alignment. *Higher Education*, Dordrecht, v. 32, p. 347-364, 1996. DOI: https://doi.org/10.1007/BF00138871. Disponível em: https://link.springer.com/article/10.1007/BF00138871. Acesso em: 15 mar. 2024.

BOLLELA, V. R.; BORGES, M. D. C.; TRONCON, L. E. A. Avaliação somativa de habilidades cognitivas: experiência envolvendo boas práticas para a elaboração de testes de múltipla escolha e a composição de exames. *Revista Brasileira de Educação Médica*, Brasília, DF, v. 42, n. 4, p. 74-85, 2018. DOI: https://doi.org/10.1590/1981-52712015v42n4rb20160065. Disponível em: https://www.scielo.br/j/rbem/a/9dnZCHRwdQKjFt7vH4DcR6n/?lang=pt. Acesso em: 30 jan. 2025.

BREDON, G. Take-home tests in economics. *Economic Analysis and Policy*, Brisbane, v. 33, n. 1, p. 52-60, 2003. DOI: https://doi.org/10.1016/S0313-5926(03)50004-2. Disponível em: https://www.sciencedirect.com/science/article/pii/S0313592603500042?via%3Dihub. Acesso em: 15 mar. 2024.

BROWN, G. T. L.; ABDULNABI, H. H. A. Evaluating the quality of higher education instructor-constructed multiple-choice tests: impact on student grades. *Frontiers in Education*, Lausanne, v. 2, 2017. DOI: https://doi.org/10.3389/feduc.2017.00024. Disponível em: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2017.00024/full. Acesso em: 30 jan. 2025.

BRYAN, C.; CLEGG, K. (ed.). *Innovative assessment in higher education*. Abingdon: Routledge, 2006.

CATANZANO, T.; JORDAN, S. G.; LEWIS, P. J. Great question! The art and science of crafting high-quality multiple-choice questions. *Journal of the American College of Radiology*, Reston, v. 19, n. 6, p. 687-692, 2022. DOI: https://doi.org/10.1016/j.jacr.2022.01.016. Disponível em: https://www.jacr.org/article/S1546-1440(22)00171-5/abstract. Acesso em: 15 mar. 2024.

COUGHLIN, P. A.; FEATHERSTONE, C. R. How to write a high quality multiple choice question (MCQ): a guide for clinicians. *European Journal of Vascular and Endovascular Surgery*, Bègles, v. 54, n. 5, p. 654-658, 2017. DOI: https://doi.org/10.1016/j.ejvs.2017.07.012. Disponível em: https://www.ejves.com/article/S1078-5884(17)30445-8/fulltext. Acesso em: 16 jun. 2024.

DELL, K. A.; WANTUCH, G. A. How-to-guide for writing multiple choice questions for the pharmacy instructor. *Currents in Pharmacy Teaching and Learning*, [*S. l.*], v. 9, n.

1, p. 137-144, 2017. DOI: https://doi.org/10.1016/j.cptl.2016.08.036. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1877129715300575?via%3Dihub. Acesso em: 30 jan. 2025.

DOWNING, S. M. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*: Theory and Practice, Dordrecht, v. 10, p. 133-143, 2005. DOI: https://doi.org/10.1007/s10459-004-4019-5. Disponível em: https://link.springer.com/article/10.1007/s10459-004-4019-5. Acesso em: 15 mar. 2024.

FAYYAZ KHAN, H.; FAROOQ DANISH, K.; SAEED AWAN, A.; ANWAR, M. Identification of technical item flaws leads to improvement of the quality of single best Multiple Choice questions. *Pakistan Journal of Medical Science*s, Karachi, v. 29, n. 3, p. 715-718, 2013. DOI: https://doi.org/10.12669/pjms.293.2993. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC3809311/. Acesso em: 30 jan. 2025.

FERNANDES, D. *Avaliação das aprendizagens*: uma agenda, muitos desafios. Lisboa: Texto Editora, 2004.

FERNANDES, D. Para um enquadramento teórico da avaliação formativa e da avaliação sumativa das aprendizagens escolares. *In*: ORTIGÃO, M. I. R.; FERNANDES, D.; PEREIRA, T. V.; SANTOS, L. (org.). *Avaliar para aprender em Portugal e no Brasil*: perspectivas teóricas, práticas e de desenvolvimento. Curitiba: CRV, 2019. p. 139-164.

FERNANDES, D.; BORRALHO, A.; BARREIRA, C.; MONTEIRO, A.; CATANI, D.; CUNHA, E.; ALVES, M. P. (ed.). *Avaliação, ensino e aprendizagem no ensino superior em Portugal e no Brasil*: realidades e perspectiva. Curitiba: EDUCA, 2014.

FERNANDES, D.; FIALHO, N. Dez anos de práticas de avaliação das aprendizagens no ensino superior: uma síntese da literatura (2000-2009). *In*: LEITE, C.; ZABALZA, M. (coord.). *Ensino superior*: inovação e qualidade na docência. Porto: Centro de Investigação e Intervenção Educativas da Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto, 2012. p. 3693-3707.

FERNANDES, S. R. G.; MACHADO, E. A.; ABELHA, M. Student assessment in higher education: a review of thesis carried out in portuguese public universities. *Revista Meta*: Avaliação, Rio de Janeiro, v. 15, n. 46, 2023. DOI: https://doi.org/10.22347/2175-2753v15i46.3844. Disponível em: https://revistas.cesgranrio.org.br/index.php/metaavaliacao/article/view/3844. Acesso em: 15 mar. 2024.

GUPTA, V.; WILLIAMS, E. R.; WADHWA, R. Multiple-choice tests: a-z in best writing practices. *Psychiatric Clinics of North America*, Filadélfia, v. 44, n. 2, p. 249-261, 2021. DOI: https://doi.org/10.1016/j.psc.2021.03.008. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0193953X21000137?via%3Dihub. Acesso em: 30 jan. 2025.

HATTIE, J. A. C.; DONOGHUE, G. M. Learning strategies: a synthesis and conceptual model. *npj Science of Learning*, Londres, v. 1, 2016. DOI: https://doi.org/10.1038/npjscilearn.2016.13. Disponível em: https://www.nature.com/articles/npjscilearn201613. Acesso em: 16 jun. 2024.

KELLAGHAN, T.; MADAUS, G. F. Outcome evaluation. *In*: STUFFLEBEAM, D. L.; MADAUS, G. F.; KELLAGHAN, T. (ed.). *Evaluation models*: viewpoints on educational and human services evaluation. Dordrecht: Springer, 2000. p. 97-112.

MORGADO, V. *et al*. Análise estrutural dos exames de ingresso nos internatos complementares de 1994 a 1996. Acta Médica *Portuguesa*, Lisboa, v. 10, n. 5, p. 387-390, 1997. DOI: https://doi.org/10.20344/amp.2427. Disponível em: https://doaj.org/article/caef5178aa5d4795b265262fbc02a0ff. Acesso em: 30 jan. 2025.

NATIONAL BOARD OF MEDICAL EXAMINERS. *Constructing written test questions for the health sciences*. 6 ed. Filadélfia: National Board of Medical Examiners, 2021.

NEDEAU-CAYO, R.; LAUGHLIN, D.; RUS, L.; HALL, J. Assessment of item-writing flaws in multiple-choice questions. *Journal for Nurses in Professional Development*, Filadélfia, v. 29, n. 2, p. 52-57, 2013. DOI: https://doi.org/10.1097/NND.0b013e318286c2f1. Disponível em: https://journals.lww.com/jnsdonline/abstract/2013/03000/assessment_of_item_writing_flaws_in.2.aspx. Acesso em: 15 mar. 2024.

PAIS, J.; SILVA, A.; GUIMARÃES, B.; POVO, A.; COELHO, E.; SILVA-PEREIRA, F.; LOURINHO, I.; FERREIRA, M. A.; SEVERO, M. Do item-writing flaws reduce examinations psychometric quality? *BMC research notes*, Londres, v. 9, 2016. DOI: https://doi.org/10.1186/s13104-016-2202-4. Disponível em: https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-016-2202-4. Acesso em: 16 jun. 2024.

SALIH, K.; JIBO, A.; ISHAQ, M.; KHAN, S.; MOHAMMED, O. A.; AL-SHAHRANI, A. M.; ABBAS, M. Psychometric analysis of multiple-choice questions in an innovative curriculum in Kingdom of Saudi Arabia. *Journal of Family Medicine and Primary Care*, Mumbai, v. 9, n. 7, p. 3663-3668, 2020. DOI: https://doi.org/10.4103/jfmpc.jfmpc_358_20. Disponível em: https://journals.lww.com/jfmpc/Fulltext/2020/09070/Psychometric_analysis_of_multiple_choice_questions.84.aspx. Acesso em: 30 jan. 2025.

SANTOS, L. Reflexões em torno da avaliação pedagógica. *In*: ORTIGÃO, M. I. R.; FERNANDES, D.; PEREIRA, T. V.; SANTOS, L. (org.). Avaliar para aprender no Brasil e em Portugal: perspectivas teóricas, práticas e de desenvolvimento. Curitiba: CRV, 2019. p. 165-190.

SANTOS, L.; PINTO, J. Ensino de conteúdos escolares: a avaliação como fator estruturante. *In*: VEIGA, F. H. (coord.). *O ensino como fator de envolvimento numa escola para todos*. Lisboa: Climepsi Editores, 2018. p. 503-539.

SCRIVEN, M. Evalution ideologies. *In*: STUFFLEBEAM, D. L.; MADAUS, G. F.; KELLAGHAN, T. (org.). *Evaluation models*: viewpoints on educational and human services evaluation. 2 ed. Dordrecht: Kluwer, 2000. p. 249-278.

SMITH, L. S. How to write better multiple-choice questions. *Nursing*, Filadélfia, v. 48, n. 11, p. 14-17, 2018. DOI: https://doi.org/10.1097/01.NURSE.0000546471.79886.85. Disponível em: https://journals.lww.com/nursing/citation/2018/11000/how_to_write_better_multiple_choice_questions.4.aspx. Acesso em: 15 mar. 2024.

TARRANT, M.; WARE, J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, [*S. l.*], v. 42, n. 2, p. 198-206, 2008. DOI: https://doi.org/10.1111/j.1365-2923.2007.02957.x. Disponível em: https://asmepublications.onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.2007.02957.x. Acesso em: 16 jun. 2024.

TAVAKOL, M.; DENNICK, R. Post-examination analysis of objective tests. *Medical Teacher*, Milton Park, v. 33, n. 6, p. 447-458, 2011. DOI: https://doi.org/10.3109/0142159X.2011.564682. Disponível em: https://www.tandfonline.com/doi/full/10.3109/0142159X.2011.564682. Acesso em: 30 jan. 2025.