

Funcionalidade dos distratores ao longo de quatro anos: Teste de Progresso do Consórcio do Centro-Oeste

Functionality of distractors over four years: Midwest Consortium Progress Test

Sulani Silva de Souza¹  sulaniamazon@gmail.com
Ana Márcia Yunes Salles Gaudard²  marciagaudard@gmail.com

RESUMO

Introdução: O teste na modalidade de múltipla escolha permite avaliar uma ampla gama de conhecimentos. Quando bem elaborado, possibilita inferências válidas sobre a capacidade dos examinandos. Os distratores – alternativas incorretas entre as opções de resposta – devem ser desafiadores e plausíveis, sendo um ponto-chave para a qualidade do teste. Este estudo teve como objetivo avaliar a funcionalidade dos distratores no Teste de Progresso ao longo de quatro anos.

Método: A escolha das opções de resposta dos estudantes do sexto ano do curso de Medicina que realizaram teste em 2017, 2018, 2019 e 2021, ou seja, teste 1, 2, 3 e 4, respectivamente, foi analisada. Classificaram-se os distratores como não funcionais quando escolhidos em menos de 5% das vezes ou não assinalados. A efetividade dos distratores foi de 0, 33,3, 66,6 e 100% para 0, 1, 2 e 3 distratores. Calcularam-se o alfa de Cronbach, a dificuldade e a discriminação. A comparação entre grupos foi realizada pela Anova.

Resultado: Considerando a ordem dos testes: 1. participaram 680, 1.006, 1.548 e 1.073 estudantes; 2. obtiveram-se os seguintes alfas de Cronbach: 0,85, 0,84, 0,82 e 0,81; 3. analisaram-se 96, 99, 100 e 98 questões, e identificaram-se 288, 297, 300 e 294 distratores; 4. dos distratores, consideraram-se funcionais 68%, 74%, 74% e 80%; 5. média e desvio padrão da dificuldade: $0,63 \pm 0,23$, $0,57 \pm 0,21$, $0,58 \pm 0,21$ e $0,54 \pm 0,23$; 6. média e desvio padrão da discriminação: $0,22 \pm 0,12$, $0,21 \pm 0,12$, $0,20 \pm 0,12$ e $0,19 \pm 0,13$; 7. média e desvio padrão dos distratores: $2 \pm 0,97$, $2,2 \pm 0,84$, $2,2 \pm 0,75$ e $2,4 \pm 0,83$. Houve diferença estatística do teste 1 para o 4 em relação à dificuldade ($p = 0,014$) e ao distrator ($p = 0,016$).

Conclusão: Houve incremento na proporção de distratores ao longo do tempo, com pico no quarto teste, que foi acompanhado de redução da dificuldade e discriminação. Uma revisão qualitativa permanente das questões se faz necessária para a adequação quanto à solidez estrutural e ao conteúdo.

Palavras-chave: Questão de Prova; Psicometria; Educação Médica; Avaliação Educacional.

ABSTRACT

Introduction: The multiple-choice test format allows the assessment of a wide range of knowledge. When well designed, it enables valid inferences about examinees' abilities. Distractors — the incorrect alternatives among the answer options — should be both challenging and plausible, as they are a key factor in test quality. The aim of this study was to evaluate the functionality of distractors in the progress test over a four-year period.

Method: The responses of sixth-year medical students who took the test in 2017, 2018, 2019, and 2021—referred to as tests 1, 2, 3, and 4, respectively—were analysed. Distractors were classified as non-functional if selected by fewer than 5% of students or not selected at all. Distractor effectiveness was categorized as 0%, 33.3%, 66.6%, or 100%, corresponding to 0, 1, 2, or 3 functional distractors per item. Cronbach's alpha, item difficulty, and item discrimination were calculated. Group comparisons were performed using ANOVA.

Results: Across the four tests: (a) 680, 1006, 1548, and 1073 students participated; (b) Cronbach's alpha values were 0.85, 0.84, 0.82, and 0.81; (c) 96, 99, 100, and 98 questions were analysed, with 288, 297, 300, and 294 distractors identified, respectively; (d) the proportion of functional distractors was 68%, 74%, 74%, and 80%; (e) the mean $\pm$ standard deviation (SD) for item difficulty was 0.63 ± 0.23 , 0.57 ± 0.21 , 0.58 ± 0.21 , and 0.54 ± 0.23 ; (f) the mean $\pm$ SD for item discrimination was 0.22 ± 0.12 , 0.21 ± 0.12 , 0.20 ± 0.12 , and 0.19 ± 0.13 ; (g) the mean $\pm$ SD for the number of functional distractors per question was 2.0 ± 0.97 , 2.2 ± 0.84 , 2.2 ± 0.75 , and 2.4 ± 0.83 . A statistically significant difference was found between tests 1 and 4 in terms of item difficulty ($p = 0.014$) and the number of functional distractors ($p = 0.016$).

Conclusion: There was an increase in the proportion of functional distractors over time, peaking in the fourth test, which was accompanied by reductions in both item difficulty and discrimination. Ongoing qualitative review of test items is essential to maintain structural integrity and content validity.

Keywords: examination questions; psychometrics; medical education; educational measurement.

¹ Universidade Católica de Brasília, Taguatinga, Distrito Federal, Brasil.

² Centro Universitário de Brasília, Brasília, Distrito Federal, Brasil.

Editora-chefe: Rosiane Viana Zuza Diniz.

Editor associado: Aristides Palhares Neto.

Recebido em 21/05/25; Aceito em 23/06/25.

Avaliado pelo processo de double blind review.

INTRODUÇÃO

O teste na modalidade de resposta escolhida permite avaliar grande abrangência de conhecimentos. Quando bem elaborado, permite inferências sobre a capacidade cognitiva dos examinandos, discriminando aqueles de alto e baixo desempenhos. As questões de múltipla escolha (QME) com resposta única consistem em uma pergunta e duas ou mais opções de respostas, e o examinando deve escolher a opção correta¹. O formato QME permite que os professores avaliem com eficiência grande número de examinados e uma ampla gama de conteúdo^{1,2}. Uma QME bem construída deve apresentar distratores eficazes. Nos processos avaliativos em que os itens do teste são projetados para medir resultados educacionais, os distratores devem ter um desempenho aceitável, e cada um deve ser baseado em um equívoco comum sobre a resposta correta, ou seja, opções não funcionais (NF) não são plausíveis nem apresentam nenhuma contribuição para a discriminação de itens³. Distratores NF são opções selecionadas com pouca frequência (< 5%) pelos examinandos ou que não apresentam o desempenho esperado. Como tal, essas opções devem ser retiradas do item ou substituídas por uma opção mais plausível⁴.

Boas práticas na construção dos itens de múltipla escolha são entendidas como características que as questões devem apresentar para que possam ser bem compreendidas por quem irá responder a elas. Trata-se de atributos de uma boa comunicação escrita, e, por isso, eles devem ser claros e objetivos. O distrator, ou seja, a opção incorreta, deve ser desafiador e plausível para que todas as opções de resposta sejam viáveis para a situação apresentada⁵. A escolha dos distratores é um ponto-chave para a qualidade do teste. É fundamental ter cuidado na sua identificação, e, para tanto, devem-se considerar os extremos, ou seja, questões muito fácil ou difícil, pois isso pode comprometer a discriminação entre os grupos e reduzir o número de opções predeterminadas para cada item apresentado e até causar confusão entre os grupos de alto e baixo desempenhos⁶. Na avaliação do conhecimento, o distrator ajuda a distinguir entre os estudantes que compreenderam o tópico e aqueles que não, uma vez que alunos que escolhem a alternativa correta podem ter um entendimento mais profundo do assunto. Distratores bem elaborados exigem que os estudantes reflitam sobre as opções antes de tomarem uma decisão. Isso promove o pensamento crítico e a análise cuidadosa das informações apresentadas. Quando bem formulados, podem desencorajar suposições aleatórias, reduzindo a descoberta da resposta correta sem compreensão do conteúdo. Eles podem oferecer *feedback* instrutivo, por apresentarem *insights* valiosos sobre os equívocos comuns dos estudantes. Isso permite que os professores identifiquem áreas que precisam ser revisadas.

As QME sempre têm algumas opções de resposta, uma das quais é a chave (resposta correta) e as demais são distrações (respostas incorretas). É importante notar que os distratores não devem ser apenas respostas erradas, mas devem ser respostas prováveis, caso o examinado cometa um erro ao procurar a opção certa. Em suma, os distratores são respostas viáveis que podem ser selecionadas quando se fazem julgamentos errados ou se verifica conhecimento (compreensão) parcial⁵.

Uma análise técnica dos itens avalia a eficiência da questão como um instrumento educativo e meio de comunicação entre avaliado e avaliador, e envolve verificar se a questão apresenta os principais “vícios” de elaboração. Os principais vícios de elaboração em relação às alternativas são: falta de sintonia gramatical entre comando da pergunta e as alternativas; falseamento de afirmações forçado pela inclusão de termos de negação como “não”, “tudo”, “todas”, “geralmente”, “apenas”, entre outros; presença de suportes com mensagens que traduzem preconceitos e aspectos negativos de duvidoso valor educativo; repetição de uma mesma forma de enunciado; inadequação do tipo de questão ao conteúdo e à habilidade; distratores obviamente errados ou sem poder para discriminar os que sabem daqueles que não sabem; presença de mais de uma alternativa correta; repetição de termos, expressões e palavras em todas as alternativas⁷.

As alternativas de resposta devem convidar o estudante a resolver o problema pela escolha da resposta correta, e, por isso, o gabarito não pode dar margem a dúvidas. Os distratores, conforme o próprio nome indica, são respostas plausíveis que têm a função de atrair quem não sabe e escolhe sem fundamento a resposta que lhe parece certa ou que lhe chama a atenção. Para evitar acertos ao acaso, os distratores têm que ser plausíveis, vale dizer, devem ser aceitáveis como possibilidades de respostas para o problema apresentado, mas não correspondem satisfatoriamente ou em sentido completo ao que é solicitado em relação ao tópico de conteúdo e à habilidade avaliados⁸.

Como redigir alternativas que sejam distratores eficientes não é fácil, os elaboradores costumam “caprichar” na opção correta e investir menos na criação dos distratores, tornando o item disfuncional. Os distratores vinculam-se ao enunciado do problema e com ele fazem sentido gramatical completo, mas são independentes entre si. Apresentam sintonia com o gabarito e são respostas corretas para outras perguntas. Mantêm uma certa similaridade de estilo e complexidade de redação com a opção correta, criando paralelismo, ou seja, um padrão homogêneo entre as alternativas de resposta por meio da sintonia entre o que todas abordam, a extensão do texto, o nível de dificuldade, a estrutura sintática, a abrangência do que enfocam, o tipo de categoria ou espécie focalizada, o estilo e a correção linguística.

Nas alternativas, não se devem misturar assuntos, criar detalhes para falsear, variar o foco, introduzir termos dúbios ou controversos, nem palavras que são pistas de resposta⁶.

Atualmente existe um conjunto de pesquisas que avaliam o número ideal de distratores em itens de múltipla escolha, mas menos pesquisas se concentraram no exame de distratores NF em QME e poucos estudos recentes examinaram especificamente a frequência de distratores NF em itens gerados pelo professor³. O número ideal de opções de resposta quando da construção de itens, algo muito discutido na literatura, pode ter apoio na dificuldade de construir os distratores adequados para cada situação apresentada.

O objetivo deste estudo é avaliar a funcionalidade dos distratores no Teste de Progresso (TP) da região Centro-Oeste comparativamente por um período de quatro anos.

MÉTODO

Analisou-se a escolha das opções de resposta dos estudantes do sexto ano do curso de Medicina que realizaram o TP de 2017, 2018, 2019 e 2021 – testes 1, 2, 3 e 4, respectivamente. Classificaram-se os distratores como NF quando escolhidos em menos de 5% das vezes ou não foram assinalados. As questões foram classificadas com 0, 1, 2 e 3 distratores, considerando que os testes utilizaram questões com quatro opções de resposta. A efetividade dos distratores (ED) foi de 0%, 33,3%, 66,6% e 100% para 0, 1, 2 e 3 distratores, respectivamente. Realizou-se o cálculo do alfa de Cronbach, da dificuldade e da discriminação das questões.

A análise de variância de uma via (Anova-One way) foi realizada com o objetivo de avaliar se havia diferenças na ED, dificuldade e discriminação das questões nos anos observados. A normalidade dos dados foi avaliada por meio dos testes Kolmogorov-Smirnov (K-S) e Shapiro-Wilk (S-W). Avaliou-se o pressuposto de homogeneidade de variância por meio do teste de Levene.

Realizaram-se procedimentos de *bootstrapping* (mil reamostragens; método de intervalo de confiança por “Bias Corrected and accelerated” - 95% IC BCa) para se obter uma maior confiabilidade dos resultados, para corrigir desvios de normalidade da distribuição da amostra e diferenças entre os tamanhos dos grupos, e, também, para apresentar um intervalo de confiança de 95% para as diferenças entre as médias⁹. O resultado foi considerado significativo quando *p* era menor que 0,05.

Esta pesquisa foi submetida ao Comitê de Ética em Pesquisa conforme Resolução nº 510/2016 do Conselho Nacional de Saúde e isenta de apreciação por analisar dados obtidos a partir de resultados das avaliações, disponibilizados por meio de plataformas institucionais de domínio público, sem qualquer possibilidade de identificação direta ou indireta dos participantes.

RESULTADO

Considerando os testes 1, 2, 3 e 4 disponíveis na Tabela 1, observaram-se os seguintes aspectos:

- Número de participantes: 680, 1.006, 1.548 e 1.073.
- Alfa de Cronbach: 0,85, 0,84, 0,82 e 0,81.
- Número de questões analisadas 96, 99, 100 e 98.
- Distratores identificados: 288, 297, 300 e 294.

Na Tabela 2, constam os seguintes valores de média e desvio padrão:

- Efetividade dos distratores: 68 ± 32, 74 ± 28, 74 ± 25 e 80 ± 28.
- Dificuldade: 0,63 ± 0,23, 0,57 ± 0,21, 0,58 ± 0,21 e 0,54 ± 0,23.
- Discriminação: 0,22 ± 0,12, 0,21 ± 0,12, 0,20 ± 0,12 e 0,19 ± 0,13.

Os testes de distribuição de normalidade de K-S e S-W demonstraram que a variável ED e a dificuldade não apresentavam distribuição normal para a maioria dos grupos

Tabela 1. Estatísticas descritivas dos testes aplicados.

TESTE	1	2	3	4	Total
Estudante (n)	680	1.006	1.548	1.073	4.307
Alfa de Cronbach	0,85	0,84	0,82	0,81	
Questões (n)	96	99	100	98	393
Distrator (n)	288	297	300	294	1179
0	8 (8%)	2 (2%)	2 (2%)	3 (3%)	15 (4%)
1	19 (20%)	20 (20%)	13 (13%)	13 (13%)	65 (16%)
2	30 (31%)	30 (30%)	45 (45%)	23 (24%)	128(33%)
3	39 (41%)	47 (47%)	40 (40%)	59 (60%)	185 (47%)

Fonte: Elaborada pelas autoras.

(ED 1, K-S (226) = 0,091; S-W (226) = 0,967; $p < 0,001$; ED 2, K-S (226) = 0,122; S-W (226) = 0,939; $p < 0,001$; ED 3, K-S (226) = 0,149; S-W (226) = 0,886; $p < 0,001$; dificuldade 1, K-S (226) = 0,091; S-W (226) = 0,967; $p < 0,001$; dificuldade 2, K-S (226) = 0,122; S-W (226) = 0,939; $p < 0,001$; dificuldade 3, K-S (226) = 0,149; S-W (226) = 0,886; $p < 0,001$; dificuldade 4, K-S (226) = 0,149;

S-W (226) = 0,886; $p < 0,001$). O teste de homogeneidade de variância de Levene demonstrou que os grupos apresentaram homogeneidade de variância para a ED, dificuldade e discriminação (Levene (3, 389) = 2,155, $p > 0,05$; Levene (3, 389) = 1,421, $p > 0,05$; Levene (3, 389) = 0,677, $p > 0,05$). Os resultados descritivos das diferenças entre os grupos encontram-se na Tabela 3.

Tabela 2. Estatísticas descritivas da efetividade dos distratores, dificuldade e discriminação para amostra total e separada por grupos.

			Estatísticas descritivas	Estimativas de <i>bootstrapping</i>	
				Intervalo de confiança da média	
				(95% IC BCa)	
				Limite inferior	Limite superior
Efetividade dos distratores	1	Média	68,03	61,64	74,79
		Desvio padrão	32,42	28,86	35,34
	2	Média	74,44	68,92	80,13
		Desvio padrão	28,18	25,23	30,61
	3	Média	74,44	69,51	79,34
		Desvio padrão	25,03	21,89	27,90
	4	Média	80,31	74,41	85,54
		Desvio padrão	27,88	23,74	31,73
	Amostra total	Média	74,34	71,25	77,42
		Desvio padrão	28,68	26,90	30,31
Dificuldade	1	Média	0,63	0,59	0,68
		Desvio padrão	0,23	0,20	0,25
	2	Média	0,58	0,54	0,62
		Desvio padrão	0,21	0,18	0,23
	3	Média	0,58	0,54	0,63
		Desvio padrão	0,21	0,18	0,23
	4	Média	0,54	0,49	0,58
		Desvio padrão	0,23	0,21	0,25
	Amostra total	Média	0,58	0,56	0,60
		Desvio padrão	0,22	0,21	0,23
Discriminação	1	Média	0,22	0,19	0,24
		Desvio padrão	0,12	0,11	0,14
	2	Média	0,21	0,19	0,24
		Desvio padrão	0,12	0,11	0,14
	3	Média	0,20	0,17	0,22
		Desvio padrão	0,12	0,10	0,13
	4	Média	0,19	0,17	0,22
		Desvio padrão	0,13	0,12	0,14
	Amostra total	Média	0,21	0,19	0,22
		Desvio padrão	0,12	0,11	0,13

Fonte: Elaborada pelas autoras.

Tabela 3. Teste *post hoc* de Hochberg com *bootstrapping* (95% IC BCa).

Comparações entre grupos			DM	Estimativas de <i>bootstrapping</i>		Tamanho de efeito (delta de Glass)
				(95% IC BCa)		
				Limite inferior	Limite superior	
<i>Efetividade dos distratores</i>	1	2	-6,42	-15,02	1,39	0,23
		3	-6,41	-14,73	2,36	0,26
		4	-12,28**	-20,27	-4,26	0,44
	2	3	0,00	-7,54	7,45	0,00
		4	-5,86	-13,41	2,18	0,21
	3	4	-5,87	-13,08	1,21	0,21
<i>Dificuldade</i>	1	2	0,06	0,00	0,12	0,28
		3	0,05	-0,01	0,11	0,24
		4	0,10**	0,03	0,16	0,41
	2	3	-0,01	-0,06	0,05	0,03
		4	0,04	-0,02	0,10	0,16
	3	4	0,04	-0,01	0,11	0,19

DM = diferença da média; ** $p < 0,05$.

Fonte: Elaborada pelas autoras.

Os resultados da Anova para a ED e dificuldade demonstraram que havia diferenças entre os grupos [$F(3, 389) = 3,010$, $p < 0,03$; $\eta^2 = 0,023$ e $F(3, 389) = 3,010$, $p < 0,03$; $\eta^2 = 0,023$]. O teste *post hoc* de Hochberg, interpretado por meio de procedimentos de *bootstrapping*, demonstrou que foram encontradas diferenças significativas entre o teste 1 e o 4 (tamanho de efeito moderado, delta de Glass = 0,44) para a ED e dificuldade (tamanho de efeito moderado, delta de Glass = 0,41). Por fim, o teste 1 não apresentou diferença significativa com teste 2 e 3 [$\Delta M = -6,42$, IC 95% BCa (-15,02 – 1,39) e $\Delta M = -6,41$, IC 95% BCa (-14,73 – 2,36)], o teste 2 não apresentou diferença significativa com teste 3 e 4 [$\Delta M = 0$, IC 95% BCa (-7,54 – 7,45) e $\Delta M = -5,86$, IC 95% BCa (-13,51 – 2,18)] e o teste 3 não apresentou diferença significativa com teste 4 [$\Delta M = -5,87$, IC 95% BCa (-13,08 – 1,21)] para a ED; o mesmo ocorreu para a dificuldade, ou seja, o teste 1 não apresentou diferença significativa com os testes 2 e 3 [$\Delta M = 0,06$, IC 95% BCa (0 – 0,12) e $\Delta M = 0,05$, IC 95% BCa (-0,01 – 0,11)], o teste 2 não apresentou diferença significativa com os testes 3 e 4 [$\Delta M = -0,01$, IC 95% BCa (-0,06 – 0,05) e $\Delta M = 0,04$, IC 95% BCa (-0,02 – 0,10)] e o teste 3 não apresentou diferença significativa com o teste 4 [$\Delta M = 0,04$, IC 95% BCa (-0,01 – 0,11)].

Os resultados da Anova para a discriminação não demonstraram que havia diferenças entre os grupos [$F(3, 389) = 3,010$, $p > 0,05$; $\eta^2 = 0,006$].

DISCUSSÃO

A avaliação do desempenho dos estudantes depende fundamentalmente da qualidade dos itens da prova. É uma

etapa fundamental no processo de análise psicométrica avaliar o desempenho dos distratores dos itens que compõem o teste como parte da documentação referente à sua confiabilidade e validade. Para realizar uma análise de distrator, é preciso coletar e analisar os dados sobre as respostas dos estudantes a cada opção, como: frequência, porcentagem e índice de dificuldade¹⁰. Além disso, deve-se comparar o desempenho dos estudantes que escolheram cada opção, observando a pontuação média, o desvio padrão e o índice de discriminação. Depois de identificar os distratores que não estão funcionando bem, como aqueles que raramente são escolhidos ou frequentemente escolhidos por estudante de alto desempenho, é possível modificá-los ou substituí-los com base nos resultados.

Como as variáveis avaliadas (ED, dificuldade e discriminação) não tiveram distribuição normal, ou seja, a curva foi assimétrica ou de livre distribuição, utilizou-se a técnica de *bootstrapping*⁹. No teste de Levene, verificou-se homogeneidade de variância nos dados dos diferentes grupos.

Na avaliação global dos testes realizada nos diferentes anos com relação à ED, as pontuações médias e o desvio padrão nos diferentes anos foram de 68 ± 32 , 74 ± 28 , 74 ± 25 e 80 ± 28 , com pontuação média foi de 74,34, o que demonstra que muitos itens tiveram três distratores funcionais, caracterizando um grande percentual de distratores funcionais comparado à literatura, com o pico no quarto teste. A alta proporção de itens com três distratores funcionais não foi totalmente surpreendente, uma vez que todos os testes foram gerados por professores universitários e submetidos a um processo de

validação por meio da análise dos itens por outro grupo de professores qualificados.

Outras pesquisas sugerem que os testes raramente têm mais de dois distratores funcionais. De acordo com Haladyna et al.⁴, aproximadamente dois terços de todos os itens de quatro opções que eles revisaram tinham apenas um ou dois distratores funcionais e nenhum dos itens de cinco opções tinha quatro distratores funcionais. Contudo, não há razão psicométrica para que todos os itens tenham o mesmo número de alternativas, uma vez algumas perguntas naturalmente teriam distratores mais ou menos plausíveis do que outras. Além disso, teste com menos opções devem, em teoria, reduzir a quantidade de tempo de leitura por item, diminuindo assim o tempo destinado para a execução do teste ou permitindo ainda que mais itens de teste sejam cobertos em uma avaliação¹¹. Portanto, embora, na maioria das circunstâncias, três alternativas sejam suficientes, os elaboradores deveriam construir os itens considerando a área de conteúdo a ser avaliada com bons distratores¹². Assim, quando se revisa um item, não devem ser eliminadas alternativas que tenham um desempenho adequado simplesmente para se conformarem a um número predefinido de opções. Ou seja, independentemente do número de opções, os itens devem ser desenvolvidos por especialistas no conteúdo, de acordo com os objetivos educacionais estabelecidos, e revisados por pares antes da aplicação para garantir que a resposta esteja inequivocamente correta e que todos os distratores sejam plausíveis¹³. No entanto, muitos testes desenvolvidos por professores, particularmente testes somativos, devem estar em conformidade com as diretrizes institucionais que definem o número de distratores por item³. Essas diretrizes raramente são baseadas em evidências e são mais propensas a ter como base as práticas de rotina e/ou procedimentos definidos, o que impede que os elaboradores tenham a flexibilidade para que possam definir os itens com variações no número de opções³. Nessas circunstâncias, dada a baixa proporção de itens com quatro distratores funcionais, itens de três opções parecem ser a escolha mais razoável. É preocupante quando é elevada a proporção de itens que não possuem distratores funcionais (12,3%). Esses itens inevitavelmente teriam estatísticas de dificuldade de item altas ($> 0,90$), caracterizando a facilidade do item, o que resulta em acertos por todos os estudantes nas instituições onde esses testes forem administrados devido a uma proporção alta de itens fáceis que provavelmente resultará na aprovação de muitos candidatos limítrofes³.

Este estudo também avaliou a confiabilidade de todos os quatro testes aplicados nos anos avaliados, os quais apresentaram valor do coeficiente alfa de Cronbach maior que 0,80, indicando uma confiabilidade satisfatória e demonstrando consistência dos resultados obtidos pelos instrumentos utilizados.

Em relação à dificuldade, a média dos índices dos quatro anos dos testes foi de 0,58; o índice de dificuldade e o desvio padrão dos testes foram: $0,63 \pm 0,23$, $0,57 \pm 0,21$, $0,58 \pm 0,21$ e $0,54 \pm 0,23$. Considerando esses resultados, fica evidente que não existem grandes diferenças em relação à dificuldade dos testes aplicados nos diferentes anos e que a dificuldade foi média. Para a dificuldade do item, as referências foram: abaixo de 0,3, difícil; entre 0,3 e 0,7, média; e acima de 0,7, fácil¹⁴. A dificuldade do item é o parâmetro mais estudado em relação ao número ideal de opções nos testes de múltipla escolha. Em metanálise realizada por Rodriguez¹⁵ concluiu-se que uma redução no número de opções sempre está associada a um aumento na dificuldade. Em um outro estudo, Tarrant et al.¹⁶ descobriram que testes com três distratores eram mais difíceis do que com quatro distratores, com dificuldade de $0,70 \pm 0,5$ versus $0,73 \pm 0,14$, respectivamente, mas os resultados não foram significativos. Muitos fatores podem ter contribuído para esse achado diferente. Por exemplo, os autores aplicaram dois testes em dois anos diferentes e com números diferentes de itens. Além disso, um número considerável de itens não era comum¹⁶.

A relação entre o número de distratores e a dificuldade do teste pode ser explicada considerando a dificuldade de os professores desenvolverem três ou mais distratores igualmente plausíveis. Um item com dois distratores plausíveis, entretanto, é preferível a um item com três ou quatro distratores implausíveis, já que os alunos raramente selecionam essas opções¹². O sucesso na elaboração dos itens está relacionado com a qualidade dos distratores e não com o número¹². Na análise do índice de discriminação das questões (correlação ponto-biserial – rpb), a média global foi de 0,21; e a discriminação dos testes de diferente ano e o desvio padrão dos testes foram: $0,22 \pm 0,12$, $0,21 \pm 0,12$, $0,20 \pm 0,12$ e $0,19 \pm 0,13$, resultados que caracterizam uma discriminação positiva.

Borgatto et al.¹⁰ relatam que o item capaz de discriminar a proficiência dos estudantes é um bom item¹⁰. Tavakol et al.¹⁷ corroboram essa ideia ao afirmarem que, quando os estudantes com melhor desempenho acertam a resposta do item e os com proficiência inferior erram o mesmo item, ele pode ser considerado adequado¹⁷. A discriminação de opções (distratores) e a discriminação de itens estão intimamente relacionadas, e é inevitável que itens com maior discriminação possuem distratores que são mais discriminatórios em geral¹⁸.

Nas análises que compararam os quatro grupos referentes aos testes realizados nos diferentes anos, encontraram-se diferenças significativas entre o teste 1 e o 4 quanto às variáveis ED e dificuldade dos itens, sem diferenças significativas nos outros grupos.

CONCLUSÃO

No presente estudo, houve um grande percentual de distratores funcionais quando comparado à literatura, com o pico no quarto teste. Houve incremento na proporção de distratores ao longo do tempo, especialmente no último teste que foi acompanhado de elevação da dificuldade do teste e diminuição da discriminação das questões. Assim, é importante um processo de qualificação das questões que envolva uma revisão permanente qualitativa delas, para a adequação quanto à solidez estrutural e do conteúdo.

Os resultados de avaliações educacionais não devem ser usados única e exclusivamente para traduzir um certo desempenho. Eles podem implicar a definição de programas de formação continuada dos docentes em que instrumentos de avaliação no ensino superior e a elaboração de itens devem ser temas abordados e discutidos.

AGRADECIMENTOS

Agradecemos a todos que, de forma direta e indireta, contribuíram para a realização deste estudo.

CONTRIBUIÇÃO DAS AUTORAS

Sulani Silva de Souza participou da concepção, análise formal, investigação e validação do estudo, e da redação do artigo. Ana Márcia Yunes Salles Gaudard participou da concepção, investigação e validação do estudo, e da redação do artigo.

CONFLITO DE INTERESSES

Declaramos não haver conflito de interesses.

FINANCIAMENTO

Declaramos não haver financiamento.

DECLARAÇÃO DE DISPONIBILIDADE DE DADOS

Os dados de pesquisa estão disponíveis no corpo do documento.
Research data is available in the body of the document.

REFERÊNCIAS

1. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*. 2004;38:327-33.
2. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach*. 2004;26:709-12.
3. Schneid SD, Armour C, Park YS, Yudkowsky R, Bordage G. Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. *Med Educ*. 2014;48(10):1020-7.
4. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? *Educ Psychol Meas*. 1993;53(4):999-1010.
5. Gierl MJ, Bulut O, Guo Q, Zhang X. Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review. *Rev Educ Res*. 2017 Dec;87(6):1082-116.
6. Instituto Nacional de Estudos e Pesquisas. Guia de elaboração e revisão de itens. Banco nacional de itens – Enade. Brasília: Inep; 2023.
7. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ*. 2002;15(3):309-34.
8. Burud I, Nagandla K, Agarwal P. Impact of distractors in item analysis of multiple-choice questions. *Int J Res Med Sci*. 2019 Mar 27;7(4):1136-9.
9. Haukoos JS, Lewis RJ. Advanced statistics: bootstrapping confidence intervals for statistics with “difficult” distributions. *Acad Emerg Med*. 2005;12(4):360-5.
10. Borgatto AF, Andrade DF. Análise clássica de testes com diferentes graus de dificuldade. *Est Aval Educ São Paulo*. 2012;52:146-56.
11. Raymond MR, Stevens C, Bucak SD. The optimal number of options for multiple-choice questions on high-stakes tests: application of a revised index for detecting nonfunctional distractors. *Adv Health Sci Educ*. 2019 Mar 1;24(1):141-50.
12. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. *Appl Meas Educ*. 1989 Jan 1;2(1):51-78.
13. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. *Adv Health Sci Educ*. 2006 Feb;11(1):61-8.
14. Pasquali L. *Psicometria: teoria dos testes na psicologia e na educação*. 4a ed. Petrópolis: Vozes; 2011.
15. Rodriguez MC. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*. 2005;24(2):1-13.
16. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med Educ*. 2009;9(1):1-8.
17. Tavakol M, Dennick R. Post-examination analysis of objective tests. *Med Teach*. 2011 June;33(6):447-58.
18. Haladyna TM, Rodriguez MC. Using full-information item analysis to improve item quality. *Educ Assess*. 2021;26(3):198-211.



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.