RBEM
Revista Brasileira de Educação Médica

ORIGINAL ARTICLE

DOI: https://doi.org/10.1590/1981-5271v49.4-2024-0143.ING

Functionality of distractors over four years: Midwest Consortium Progress Test

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

Sulani Silva de Souza[1] sulaniamazon@gmail.com
Ana Márcia Yunes Salles Gaudard[2] marciagaudard@gmail.com

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 ± 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 ± 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 ± 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.

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 ± 0, 23, 0,57 ± 0,21, 0,58 ± 0,21 e 0,54 ± 0,23; 6. média e desvio padrão da discriminação: 0,22 ± 0,12, 0,21 ± 0,12, 0,20 ± 0,12 e 0,19 ± 0,13; 7. média e desvio padrão dos distratores: 2 ± 0,97, 2,2 ± 0,84, 2,2 ± 0,75 e 2,4 ± 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.*

[1] Universidade Católica de Brasília, Taguatinga, Distrito Federal, Brazil.
[2] Centro Universitário de Brasília, Brasilia, Distrito Federal, Brazil.

Chief Editor: Rosiane Viana Zuza Diniz.
Associate Editor: ARISTIDES PALHARES NETO.

Received on 05/21/25; Accepted on 06/23/25.

Evaluated by double blind review process.

INTRODUCTION

The test in the chosen response modality allows the evaluation of a wide range of knowledge. When well designed, it allows inferences about the cognitive capacity of the examinees, discriminating between those with high and low performance. Multiple choice questions (MCQ) with a single answer consist of one question, two or more answer options in which the examinee must choose the correct option [1]. The MCQ format allows teachers to efficiently assess large numbers of examinees and a wide range of content[1,2]. A well-constructed MCQ should feature effective distractors. In evaluation processes where test items are designed to measure educational outcomes, distractors must perform acceptably and each should be based on a common misconception about the correct answer, i.e., non-functional options are not plausible or have any contribution to item discrimination[3]. Non-functional distractors are options that are infrequently selected (< 5%) by examinees or that do not perform as expected. As such, these options should be removed from the item or replaced with a more plausible option [4].

Good practices in the construction of multiple-choice items are understood as characteristics that must be presented to be well understood by those who will answer it. These are attributes of good written communication and include, above all, clarity and objectivity. The distractor, that is, the incorrect option, must be challenging and plausible so that all response options are viable for the situation presented [5]. The choice of distractors is a key point for the quality of the test. It is essential to be careful in its identification considering the extremes, that is, very easy or difficult questions, as it can compromise discrimination between groups and reduce the number of predetermined options for each item presented and even cause confusion between high and low performance groups [6]. When assessing knowledge, it helps to distinguish between students who have understood the topic and those who have not, since students who choose the correct alternative can have a deeper understanding of the subject. Well-designed distractors require students to think about options before deciding. This promotes critical thinking and careful analysis of the information presented. When well formulated they can discourage random assumptions, reducing the discovery of the correct answer without understanding the content. They can offer instructive feedback by providing valuable insights into common student misconceptions. This allows teachers to identify areas that need to be reviewed.

Multiple-choice questions always have a few answer options, one of which is the key (correct answer) and the rest are distractions (incorrect answers). It is important to note that distractors should not only have wrong answers, but should be probable answers, in case the examinee makes a mistake when looking for the right option. In short, distractors are viable responses that one can select by making wrong judgments or having partial knowledge (understanding).[5]

A technical analysis of the items evaluates the efficiency of the question as an educational instrument and means of communication between the evaluated and the evaluator and involves verifying whether the question has the main "defects" of creation. The main creation defects in relation to the alternatives are: Lack of grammatical harmony between the command of the question and the alternatives; Falsification of statements forced by the inclusion of negation terms such as "no", "everything", "all", "generally", "only", among others; Presence of supports with messages that translate prejudices and negative aspects of dubious educational value; Repetition of the same form of statement; Inadequacy of the type of question to the content and ability; Distractors which are obviously wrong or powerless to discriminate between those who know and those who don't; Presence of more than one correct alternative; Repetition of terms, expressions and words in all alternatives [7].

The answer alternatives must invite the student to solve the problem by choosing the correct answer, so the answer key cannot leave room for doubt. Distractors, as their name implies, are plausible answers that have the function of attracting those who do not know and choose without foundation the answer that seems right or that catches their attention. To avoid random hits, the distractors must be plausible, that is, they must be acceptable as possible answers to the problem presented, but they do not correspond satisfactorily or completely to what is requested in relation to the topic of content and skill evaluated [8].

Writing alternatives that are efficient distractors is not easy, so developers tend to "be meticulous" with the correct option and invest less in the creation of distractors, making the item dysfunctional. The distractors are linked to the statement of the problem and with it make complete grammatical sense, but they are independent of each other. They are in tune with the answer key and are correct answers to other questions. They maintain a certain similarity of style and complexity of writing with the correct option, creating parallelism, that is, a homogeneous pattern between the answer alternatives through the harmony between what they all address, the length of the text, the level of difficulty, the syntactic structure, the scope of what they focus on, the type of focused category or species, style and linguistic correctness. In the alternatives, one should not mix subjects, create details to falsify, vary the focus, introduce dubious or controversial terms, or words that are clues to answer [6].

There is currently a body of research evaluating the optimal number of distractors in multiple-choice items,

but less research has focused on the examination of non-functional distractors in MCQ, and few recent studies have specifically examined the frequency of non-functional distractors in teacher-generated items[3]. The ideal number of response options when constructing items, something much discussed in the literature, may be supported by the difficulty of constructing the appropriate distractors for each situation presented.

The objective of this study is to evaluate the functionality of distractors in the progress test (PT) of the Midwest region comparatively over a period of four years.

METHOD

The choice of response options of sixth-year medical students who took the 2017, 2018, 2019 and 2021 PT called tests 1, 2, 3 and 4 respectively were analyzed. Distractors were classified as non-functional (NF) when chosen less than 5% of the time or were not chosen. The questions were classified with 0, 1, 2 and 3 distractors, considering that the tests used questions with four answer options. The effectiveness of distractors (ED) was 0, 33.3, 66.6 and 100% for 0, 1, 2 and 3 distractors, respectively. Cronbach's alpha, difficulty and discrimination of the questions were calculated.

One-way analysis of variance (ANOVA-One way) was performed with the objective of evaluating whether there were differences in ED, difficulty and discrimination of the questions in the observed years. The normality of the data was evaluated using the Kolmogorov-Smirnov (K-S) and Shapiro-Wilk (S-W) tests. The assumption of homogeneity of variance was evaluated using the Levene test.

Bootstrapping procedures (1000 re-sampling; confidence interval method by bias corrected and accelerated - 95%CI BCa) were performed to obtain greater reliability of the results, to correct deviations from normality in the sample distribution, differences between group sizes, and to present a 95% confidence interval for differences between means[9]. The result was considered significant when p was less than 0.05.

This research was submitted to the Research Ethics Committee according to resolution Nº. 510/2016 of the National Health Council and was exempt from appreciation for analyzing data obtained from the results of the evaluations, made available through institutional platforms in the public domain, without any possibility of direct or indirect participants' identification.

RESULTS

Considering tests 1, 2, 3 and 4 in Table 1 it was observed: a) Number of participants: 680, 1006, 1548 and 1073; b) Cronbach's alpha: 0.85, 0.84, 0.82 and 0.81; c) Number of questions analyzed 96, 99, 100 and 98 and d) Identified distractors 288, 297, 300 and 294 and in Table 2 the mean and standard deviation for: a) Effectiveness of the distractors: 68 ± 32, 74 ± 28, 74 ± 25 and 80 ± 28; b) Difficulty: 0.63 ± 0.23, 0.57 ± 0.21, 0.58 ± 0.21 and 0.54 ± 0.23; c) Discrimination: 0.22 ± 0.12, 0.21 ± 0.12, 0.20 ± 0.12 and 0.19 ± 0.13.

Tests of the distribution of normality of K-S and S-W showed that the variable ED and difficulty did not have a normal distribution for most groups (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; $p < 0.001$; ED 4, K-S (226) = 0.149; S-W (226) = 0.886; $p < 0.001$; Difficulty 1, K-S (226) = 0.091; S-W (226) = 0.967; $p < 0.001$; Difficulty 2, K-S (226) = 0.122; S-W (226) = 0.939; $p < 0.001$; Difficulty 3, K-S (226) = 0.149; S-W (226) = 0.886; $p < 0.001$; Difficulty 4, K-S (226) = 0.149; S-W (226) = 0.886; $p < 0.001$). Levene's test of homogeneity of variance showed that the groups showed homogeneity of variance for ED, difficulty and discrimination (Levene (3, 389) = 2.155, $p > 0.05$; Levene (3, 389) = 1.421, $p > 0.05$; Levene (3, 389) = 0.677, $p > 0.05$). The descriptive results of the differences between the groups are shown in Table 3.

Table 1. Descriptive statistics of the applied tests:

TEST	1	2	3	4	Total
Student (n)	680	1006	1548	1073	4307
Cronbach's *alpha*	0.85	0.84	0.82	0.81	
Questions (n)	96	99	100	98	393
Distractor (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%)

Source: prepared by the authors.

 DOI: https://doi.org/10.1590/1981-5271v49.4-2024-0143.ING

Table 2. Descriptive statistics of the effectiveness of the distractors, difficulty and discrimination for the total sample and separated by groups.

				Bootstrapping Estimates	
			Descriptive Statistics	Mean Confidence Interval (95% IC BCa)	
				Lower Limit	Upper Limit
Effectiveness of distractors	1	Mean	68.03	61.64	74.79
		Standard deviation	32.42	28.86	35.34
	2	Mean	74.44	68.92	80.13
		Standard deviation	28.18	25.23	30.61
	3	Mean	74.44	69.51	79.34
		Standard deviation	25.03	21.89	27.90
	4	Mean	80.31	74.41	85.54
		Standard deviation	27.88	23.74	31.73
	Total Sample	Mean	74.34	71.25	77.42
		Standard deviation	28.68	26.90	30.31
Difficulty	1	Mean	0.63	0.59	0.68
		Standard deviation	0.23	0.20	0.25
	2	Mean	0.58	0.54	0.62
		Standard deviation	0.21	0.18	0.23
	3	Mean	0.58	0.54	0.63
		Standard deviation	0.21	0.18	0.23
	4	Mean	0.54	0.49	0.58
		Standard deviation	0.23	0.21	0.25
	Total Sample	Mean	0.58	0.56	0.60
		Standard deviation	0.22	0.21	0.23
Discrimination	1	Mean	0.22	0.19	0.24
		Standard deviation	0.12	0.11	0.14
	2	Mean	0.21	0.19	0.24
		Standard deviation	0.12	0.11	0.14
	3	Mean	0.20	0.17	0.22
		Standard deviation	0.12	0.10	0.13
	4	Mean	0.19	0.17	0.22
		Standard deviation	0.13	0.12	0.14
	Total Sample	Mean	0.21	0.19	0.22
		Standard deviation	0.12	0.11	0.13

Source: prepared by the authors.

Table 3. Post-hoc Hochberg test with Bootstrapping (95% IC BCa).

	Comparisons between groups		DM	Bootstrapping Estimates (95% CI BCa)		Effect Size (Glass' Delta)
				Lower limit	Upper Limit	
Effectiveness of distractors	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
Difficulty	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 = difference from the mean; ** $p < 0.05$.
Source: prepared by the authors.

The ANOVA results for ED and difficulty showed that there were differences between the groups [F (3, 389) = 3.010, $p < 0.03$; $\eta 2 = 0.023$ and F (3, 389) = 3.010, $p < 0.03$; $\eta 2 = 0.023$]. Hochberg's post-hoc test, interpreted by means of bootstrapping procedures, showed that significant differences were found between tests 1 and 4 (moderate effect size, Glass' delta = 0.44) for ED and difficulty (moderate effect size, Glass' delta = 0.41). Finally, test 1 did not show any significant difference with tests 2 and 3 [ΔM = -6.42, 95% CI BCa (-15.02 – 1.39) and ΔM = -6.41, 95% CI BCa (-14.73 – 2.36)], test 2 did not show any significant difference with tests 3 and 4 [(ΔM = 0, 95% CI BCa (-7.54 – 7.45) and ΔM = -5.86, 95% CI BCa (-13.51 – 2.18)] and test 3 did not show any significant difference with test 4 [(ΔM = -5.87, 95% BCa CI (-13.08 – 1.21)] for ED; the same occurred for difficulty, i.e., test 1 did not show any significant difference with tests 2 and 3 [(ΔM = 0.06, 95% CI BCa (0 – 0.12) and ΔM = 0.05, 95% CI BCa (-0.01 – 0.11)], test 2 did not show any significant difference with tests 3 and 4 [(ΔM = -0.01, 95% CI BCa (-0.06 – 0.05) and ΔM = 0.04, 95% CI BCa (-0.02 – 0.10)] and test 3 showed no significant difference with test 4 [(ΔM = 0.04, 95% CI BCa (-0.01 – 0.11)].

The ANOVA results for discrimination did not demonstrate that there were differences between the groups [F (3, 389) = 3.010, $p > 0.05$; $\eta 2 = 0.006$].

DISCUSSION

The evaluation of student performance depends fundamentally on the quality of the test items. It is a fundamental step in the psychometric analysis process to evaluate the performance of the distractors of the items that constitute the test as part of the documentation regarding its reliability and validity. To perform a distractor analysis, it is necessary to collect and analyze data on the students' responses to each option, such as frequency, percentage, and difficulty index[10]. In addition, it must compare the performance of the students who chose each option, observing the average score, the standard deviation and the discrimination index. Once you have identified the distractors that are not working well, such as those that are rarely chosen or often chosen by high-achieving students, you can modify or replace them based on the results.

As the variables evaluated (ED, difficulty and discrimination) did not have normal distribution, that is, the curve was asymmetric or freely distributed, the bootstrapping[9] technique was used. In Levene's test, homogeneity of variance was verified in the data of the different groups.

In the global evaluation of the tests in the different years in relation to ED, the mean scores and standard deviation in the different years were 68 ± 32, 74 ± 28, 74 ± 25 and 80 ± 28, and the mean score was 74.34, which demonstrates that many items had three functional distractors, characterizing a large percentage of functional distractors compared to the literature, with the peak in the fourth test. The high proportion of items with three functional distractors was not entirely surprising, since all tests were generated by university professors and underwent a validation process through the analysis of the items by another group of qualified teachers.

Other studies suggest that tests rarely have more than two functional distractors. Haladyna et al found that

approximately two-thirds of all of the four-option items they reviewed had only one or two functional distractors and none of the five-option items had four functional distractors [4]. On the other hand, there is no psychometric reason why all items should have the same number of alternatives, since some questions would naturally have more or less plausible distractors than others. In addition, tests with fewer options should, in theory, reduce the amount of reading time per item, thereby decreasing the time allotted for test performance or allowing even more test items to be covered in an assessment [11]. Therefore, although in most circumstances, three alternatives would be sufficient, the creators should construct the items considering the content area to be evaluated with good distractors[12]. Thus, when reviewing an item, alternatives that perform adequately simply to conform to a predefined number of options should not be eliminated. That is, regardless of the number of options, the items must be developed by experts in the content, according to the established educational objectives, and reviewed by peers before application to ensure that the answer is unequivocally correct and that all distractors are plausible[13]. However, many teacher-developed tests, particularly summative tests, must conform to institutional guidelines that define the number of distractors per item [3]. These guidelines are rarely evidence-based and are more likely to be based on routine practices and/or defined procedures, preventing creators from having the flexibility to define items with variations in the number of options[3]. In these circumstances, given the low proportion of items with four functional distractors, three-choice items seem to be the most reasonable option. It is a matter of concern whether the proportion of items that do not have functional distractors is high (12.3%). These items would inevitably have high item difficulty statistics (> 0.90) characterizing the facility of the item, resulting in correct answers by all students at the institutions where these tests are administered due to a high proportion of easy items that will likely result in the approval of many borderline candidates [3].

This study also evaluated the reliability of all four tests applied in the years evaluated, and these presented Cronbach's alpha coefficient value greater than 0.80, indicating satisfactory reliability and demonstrating consistency of the results obtained by the instruments used.

Regarding difficulty, the mean index of the four years of the tests was 0.58; the difficulty index and the standard deviation of the tests were: 0.63 ± 0.23, 0.57 ± 0.21, 0.58 ± 0.21 and 0.54 ± 0.23. Taking these results into account, it is evident that there are no major differences in relation to the difficulty of the tests applied in the different years and the difficulty was medium. For the difficulty of the item, the references were: below 0.3, difficult; between 0.3 and 0.7, medium and above 0.7, easy [14]. The difficulty of the item is the most studied parameter in relation to the ideal number of options in the multiple-choice tests. In a meta-analysis carried out by Rodriguez, it was concluded that a reduction in the number of options is always associated with an increase in difficulty[15]. In another study, by Tarrant et al, they found that tests with three distractors were more difficult than with four distractors, with a difficulty of 0.70 ± 0.5 versus 0.73 ± 0.14, respectively, but the results were not significant[16]. Many factors may have contributed to this different finding. For example, the authors applied two tests in two different years and with different numbers of items. In addition, a considerable number of items were not common [16].

The relationship between the number of distractors and the difficulty of the test can be explained by considering the difficulty of teachers developing three or more equally plausible distractors. One item with two plausible distractors, however, is preferable to an item with three or four implausible distractors, since students rarely select these options [12]. Success in the creation of the items is related to the quality of the distractors and not to the number [12]. In the analysis of the discrimination index of the questions (point-biserial correlation - rpb), the global mean was 0.21; and the discrimination of the tests of different years and the standard deviation of the tests were 0.22 ± 0.12, 0.21 ± 0.12, 0.20 ± 0.12 and 0.19 ± 0.13, results that characterize positive discrimination.

Borgatto et al report that the item capable of discriminating the students' proficiency is a good item [10]. Tavakol et al corroborate this idea by stating that when students with better performance get the correct answer to the item and those with lower proficiency get the same item wrong, it can be considered adequate[17]. Option discrimination (distractors) and item discrimination are closely related, and it is inevitable that items with greater discrimination have distractors that are more discriminatory overall [18].

In the analysis comparing the four groups related to the tests performed in the different years, it was demonstrated that significant differences were found between tests 1 and 4 regarding the variables ED and difficulty of the items, with no significant differences in the other groups.

CONCLUSION

In the present study, there was a large percentage of functional distractors compared to literature, with the peak in the fourth test. There was an increase in the proportion of distractors over time, especially in the last test, which was accompanied by an increase in the difficulty of the test and a decrease in the discrimination of the questions. Thus, it is important to have a process of qualification of the questions

that involves a permanent qualitative review of the questions, for adequacy as to the structural and content soundness.

The results of educational evaluations should not be used solely and exclusively to translate a certain performance. They may imply the definition of continuing education programs for teachers, where evaluation instruments in higher education and the creation of items should be an addressed and discussed topic.

ACKNOWLEDGMENTS

To all those who directly and indirectly collaborated with this study.

AUTHORS' CONTRIBUTIONS

Sulani Silva de Souza contributed to the conception, formal analysis, investigation, validation and writing of the manuscript. Ana Márcia Yunes Salles Gaudard contributed to the conception, research, validation and writing of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

SOURCES OF FUNDING

The authors declare no sources of funding.

DATA AVAILABILITY STATEMENT

Research data is available in the body of the document

REFERENCES

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 1º;24(1):141-50.
12. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. Appl Meas Educ. 1989 Jan 1º;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.