·教育评价·

医学考试中减少选择题无效干扰选项方法研究

李国建[1]　卢燕[2]　何惧[3]　刘茂伟[1]　冯攀[4]
[1]国家医学考试中心试题开发一部，北京　100097；[2]国家医学考试中心发展研究部，北京　100097；[3]国家医学考试中心，北京　100097；[4]国家医学考试中心综合管理部，北京　100097
通信作者：卢燕，Email：mingtiyichu@163.com

【摘要】 目的　观察无效干扰选项修改后试题难度、区分度及被修改选项选择率变化，探讨通过修改无效干扰选项提高试题质量的可行性。**方法**　以某医学基本知识考试中包含无效干扰选项的试题32题为研究对象。请专家根据命审题工作要求，在实测数据提示下，对无效干扰选项进行分析修改，观察修改前后试题的难度、区分度变化及其相关性和被修改选项选择率变化。采用SPSS 24.0进行数据处理。试题修改前后的难度、区分度及修改选项的选择率变化采用配对Wilcoxon检验；修改前和修改后试题难度、区分度相关性采用斯皮尔曼等级相关法。**结果**　少数试题（6题）修改原因为技术性因素，多数试题（26题）修改原因为知识内容方面干扰性不强。对一个无效干扰选项进行修改后，修改前后试题难度分别为0.676（0.558，0.893）和0.637（0.531，0.839），区分度（点二列相关）分别为0.261（0.150，0.316）和0.262（0.138，0.358），区分度（鉴别指数）分别为0.215（0.113，0.352）和0.259（0.138，0.346）。修改前后难度和鉴别指数差异有统计学意义。修改前后难度和区分度均相关。被修改选项的选择率修改前为0.009（0.003，0.015），修改后为0.044（0.021，0.092），差异有统计学意义。**结论**　本研究涉及的无效干扰选项出现的原因多数为试题内容方面，技术性缺陷相对较少。通过对无效干扰选项的分析和修改，可以有效提高无效干扰选项的选择率，提高试题难度和区分度。

【关键词】 医学考试；　选择题；　干扰选项；　难度；　区分度
【中图分类号】 R-05
DOI：10.3760/cma.j.cn116021-20230509-01565

A study on reducing nonfunctional distractors of multiple-choice questions in medical examination

Li Guojian[1], Lu Yan[2], He Ju[3], Liu Maowei[1], Feng Pan[4]
[1]Item Development Division 1, National Medical Examination Center, Beijing 100097, China; [2]Research and Development Department, National Medical Examination Center, Beijing 100097, China; [3]National Medical Examination Center, Beijing 100097, China; [4]General Management Department, National Medical Examination Center, Beijing 100097, China
Corresponding author: Lu Yan, Email: mingtiyichu@163.com

【Abstract】 Objective　To discuss the possibility of modifying nonfunctional distractors to improve the quality of multiple-choice questions in medical examination through analyzing the changes in the difficulty and discrimination of items and the selection rate of modified distractors. **Methods**　Thirty-two multiple-choice questions involving nonfunctional distractors from a medical examination were studied. According to the item-writing guidelines, experts modified nonfunctional distractors based on actual measurements. We analyzed the changes in the difficulty and discrimination of items and the selection rate of modified distractors before and after item modification using the paired Wilcoxon test, and investigated the correlation of item difficulty and discrimination before and after item modification by Spearman's rank correlation analysis. **Results**　Six questions were modified due to technical factors, and 26 questions were

modified due to weak interference in knowledge content. Before and after nonfunctional distractor modification, the difficulty values of the questions were 0.676 (0.558, 0.893) and 0.637 (0.531, 0.839), respectively; the discrimination values (point-biserial correlation) were 0.261 (0.150, 0.316) and 0.262 (0.138, 0.358), respectively; and the discrimination index values were 0.215 (0.113, 0.352) and 0.259 (0.138, 0.346), respectively. There were significant differences in the difficulty and discrimination index of items before and after modification. The difficulty and discrimination of items before and after modification were both significantly correlated. The selection rates of modified distractors were 0.009 (0.003, 0.015) and 0.044 (0.021, 0.092) before and after modification, respectively, which were significantly different. **Conclusions** Nonfunctional distractors in this study were mainly caused by reasons in the contents of the test questions, less in technical defects. Through the analysis and modification of nonfunctional distractors, the selection rate of nonfunctional distractors can be effectively increased, and item difficulty and discrimination can be improved.

【Key words】 Medical examination; Multiple-choice question; Distractor; Difficulty; Discrimination

DOI: 10.3760/cma.j.cn116021-20230509-01565

1　问题提出

1.1　选择题历史悠久，是考试常用题型

选择题诞生一百余年，特别是计算机化考试实现以来，被认为是目前最经济、有效、持久的教育评估形式之一[1]。选择题由题干、选项和辅助信息组成。选项为一组备选答案，其中包括一个正确选项和一个或多个错误选项（也称“干扰选项”）。干扰选项的相对正确性不同，可能完全错误或者部分错误，学生必须运用知识和解决问题的能力来确定题干内容对应的正确选项，而干扰选项功能差异可能影响学生作答和考试质量[2-5]。

1.2　无效干扰选项在考试中普遍存在

一般将选择率小于5%和/或区分度大于零的干扰选项定义为无效干扰选项[6-8]。国内外多项研究文献报告，医学考试中无效干扰选项在所有干扰选项中所占比例为47.80%[9]~68.30%[6]（Mark为58.90%[10]，Kilgour和Tayyaba为55.30%[11]）。

由于大多数教育考试和资格认证考试被设计成标准参照考试，为了满足考试设计和考试内容覆盖要求，测试中不可避免存在相当数量的简单试题。简单试题正确答案选择率高，干扰选项选择率必然低。因此，传统无效干扰选项定义在一定程度上不适用于这类考试。近期有学者提出修正的无效干扰选项的计算方法[10]，即$p_{nfd}=0.1-(p_c\times 0.1)$，其中$p_{nfd}$（nonfunctional distractor）表示某一试题中确定无效干扰选项标准，p_c为试题难度值。据此修正公式，不同难度试题的无效干扰选项的界定不同。在试题难度为0.5的情况下，符合选择率小于0.05为无效干扰选项的传统方法；当试题难度大于0.5，无效干扰选项选择率界定减小；当试题难度小于0.5，则无效干扰选项选择率界定增大。该研究统计了修正标准下某次平均难度为0.743的考试中无效干扰选项比例为34.70%。

可见不论以传统标准还是修正标准进行无效干扰选项的判别，无效干扰选项的普遍存在都是客观事实。无效干扰选项的存在，必然影响试题难度和试题区分度，从而影响考试质量。减少或尽可能避免试题中出现无效干扰选项，是提高试题质量的要求之一。

1.3　五选项最佳选择题为医学考试常用题型

五选项最佳选择题是医学教学和考试中常用的题型之一。它包括1个正确选项和4个干扰选项，在医学教育考核和执业准入中普遍应用。目前，我国医师资格考试医学综合笔试题型为五选项选择题。

虽然相关研究认为将干扰选项减少到2~3个不会对考试质量造成影响，并有利于提高命题和考试效率[8]。但也有学者认为医学考试选择题在每道试题至少设置2个有效干扰选项的基础上，在确保质量的前提下，应尽量多地开发有效干扰选项[12]。

1.4　研究无效干扰选项修改效果具有一定的现实意义

随着医学考试改革的深化，在一年多考、突发情况部分地区补考等复杂形势下，考试对高质量试题的需求量越来越大，对准确把握试题难度和区分

度的要求越来越高。请专家深入分析题库中现有试题干扰选项在实际测量中的具体表现，在不改变试题题干、正确选项（即标答）和干扰选项数量的情况下，对无效干扰选项进行修改并观察其选择率及试题难度变化，可以使专家重视干扰项对试题质量的影响，提高专家提高试题质量、控制命题难度的能力。在题库建设方面，对固定考点试题，可通过调整干扰选项的内容形成不同难度的试题，以满足不同组卷需求。

2　研究方法

以某医学基本知识考试中涉及的全部无效干扰选项中选择率最低的无效干扰选项的32道题为研究对象。对选择率最低的一个干扰选项进行修改，并在第二年考试中再次施测，比较修改前和修改后试题难度、区分度变化及相关性，以及被修改选项选择率变化。

2.1　样本选择

本研究涉及某医学考试试卷总题量为330题，考生来自参加当年考试的全部24所院校的所有五年制四年末考生，考生人数为7 330人。根据实测数据，从全部1 320个干扰选项中选择包含选择率最低的干扰选项的32道试题（每道试题可能含有多个无效干扰选项），请命题专家按命题要求对每道试题中选择率最低的一个无效干扰选项进行修改，题干和其他选项保持不变。修改后的试题在第二年再次施测，测试对象为相同24所院校五年制四年末全部考生共10 791人，并通过共同题设计认为相邻年度考生水平基本一致。

32道试题中A1型试题16道、A2型试题16道；覆盖全部4个模块，其中临床医学19题、基础医学8题、预防医学1题、医学人文4题；记忆11题、理解6题、应用15题（表1）。

2.2　样本处理

依据试题开发的要求，干扰选项应在性质上和类别上相同，在逻辑、语法结构和内容上长短基本一致，要避免相互重叠、依赖，避免无意义干扰选项，不能使用“以上都是”或“以上都不是”，并与题干存在逻辑关系，这与国际通用做法相符[1]。依据以上要求再次审视32个无效干扰选项，分析其选择率低的可能原因，并进行修改。处理后的试题在第二年再次施测。

表1　研究试题参数

序号	题型	专业	认知层次	序号	题型	专业	认知层次
1	A1	输血	理解	17	A2	泌外	简单应用
2	A1	生化	记忆	18	A2	呼吸	简单应用
3	A1	药理	记忆	19	A2	法规	综合应用
4	A1	微生物	记忆	20	A2	预防	综合应用
5	A1	消化	记忆	21	A2	妇科	简单应用
6	A1	解剖	记忆	22	A2	中毒	综合应用
7	A1	儿科	理解	23	A2	骨科	理解
8	A1	病理	记忆	24	A2	普外	简单应用
9	A1	生理	理解	25	A2	伦理	综合应用
10	A1	伦理	理解	26	A2	精神	简单应用
11	A1	免疫	记忆	27	A2	风湿	简单应用
12	A1	心血管	记忆	28	A2	内分泌	简单应用
13	A1	产科	理解	29	A2	外总	简单应用
14	A1	神内	记忆	30	A2	儿科	简单应用
15	A1	免疫	记忆	31	A2	普外	简单应用
16	A1	心理	记忆	32	A2	产科	简单应用

2.3　统计学方法

以SPSS 24.0对数据进行统计学处理。因试题的难度、区分度及选择率为相对数，一般不服从正态分布，因此用中位数和四分位数表示。采用非参数的检验方法（配对Wilcoxon检验）比较32道试题修改前后的难度、区分度及修改选项的选择率变化；采用斯皮尔曼等级相关计算修改前和修改后试题难度、区分度的相关性。其中试题难度以通过率表示，即答对该试题的人数占考生总人数的百分比，计算公式为$P=R/N$，其中P代表试题的答对率，R为答对该试题的人数，N为全体考生人数。P值越大，答对人数比例越高，试题越容易。试题区分度采用相关系数法和鉴别指数法。相关系数法指试题得分与试卷总分的点二列相关系数，也称为点二列相关法。鉴别指数法计算是将全体考生总分从高到低排列，将总分最高的27.00%考生定为高分组，总分最低的27.00%考生定为低分组，分别计算两组考生在某道试题上的答对率，两个答对率之差就是这道题的区分度。区分度反映试题对考生心理特征的区分能力，区分度高的试题能将不同水平的考生区分开来，一般认为0.300以上表明试题区分度较好。

3　结果

第一年全卷难度为0.628，无效干扰选项比例按

传统定义为44.62%，按修正标准判断为27.73%。选出的32题无效干扰选项比例按两种标准判断分别为63.28%和36.72%。第二年全卷难度为0.614，无效干扰选项比例按传统定义为46.29%，按修正标准判断为28.71%。修改的32道题无效干扰选项比例按两种标准判断分别为50.00%和30.47%。与国内外相关报道相比，本次研究涉及的两套中等难度试卷无效干扰选项的比例低于文献报道，选出的32道题中无效干扰选项比例与文献报道数据持平。

无效干扰选项修改原因：多数（26题）为知识内容方面干扰性不强；少数（6题）为技术方面问题，具体为选项中含有不专业词汇3题，选项表述过于绝对2题，干扰选项与其他选项同质性差1题。

3.1 修改前后试题难度和区分度变化

修改前后试题难度的中位数和四分位数$M(Q_1, Q_3)$为0.676（0.558，0.893）和0.637（0.531，0.839）；区分度（点二列相关）$M(Q_1, Q_3)$为0.261（0.150，0.316）和0.262（0.138，0.358）；区分度（鉴别指数）$M(Q_1, Q_3)$为0.215（0.113，0.352）和0.259（0.138，0.346）。配对Wilcoxon检验结果显示，修改前后难度及区分度（鉴别指数）差异有统计学意义，区分度（点二列相关）差异无统计学意义（表2）。通过修改无效干扰选项，可以提高试题区分度（鉴别指数），也不可避免地提高了试题难度。

表2 修改前后试题难度、区分度变化（n=32）

指标		$M(Q_1, Q_3)$	Z值	P值
难度	修改前	0.676（0.558，0.893）	−4.04	<0.001
	修改后	0.637（0.531，0.839）		
区分度（点二列相关）	修改前	0.261（0.150，0.316）	−1.85	0.064
	修改后	0.262（0.138，0.358）		
区分度（鉴别指数）	修改前	0.215（0.113，0.352）	−2.84	0.004
	修改后	0.259（0.138，0.346）		

3.2 修改前后试题难度和区分度相关性

修改前后试题难度斯皮尔曼相关系数r=0.944，修改前后试题区分度（点二列相关）斯皮尔曼相关系数r=0.904，区分度（鉴别指数）斯皮尔曼相关系数r=0.889，统计检验均显著，呈高度相关（表3）。

表3 修改前后试题难度和区分度之间的相关性（n=32）

指标	相关系数	P值
难度	0.944	<0.001
区分度（点二列相关）	0.904	<0.001
区分度（鉴别指数）	0.889	<0.001

3.3 被修改选项选择率变化

被修改选项选择率$M(Q_1, Q_3)$从0.009（0.003，0.015）变为修改后的0.044（0.021，0.092），采用非参数的检验方法（配对Wilcoxon检验），结果显示差异有统计学意义（表4）。无效干扰选项通过专家修改后，可以大幅度提高其选择率，增加干扰性。

表4 被修改选项选择率变化的非参数检验结果

指标		$M(Q_1, Q_3)$	Z值	P值
选择率	修改前	0.009（0.003，0.015）	−4.88	<0.001
	修改后	0.044（0.021，0.092）		

4 讨论

4.1 遵从试题开发程序可维持无效干扰选项比例在同行业的相当水平

在命审题过程中，干扰选项的设置相对于试题内容和标准答案更容易被忽略，虽然在命审题环节严格遵从特定原则对试题进行过多轮反复审校，但不可避免存在某些干扰选项不完美的情况。本研究涉及的考试连续两年无效干扰选项比例为44.62%和46.29%，低于目前文献报道的47.80%[9]~68.30%[6]，按照修正标准分别为27.73%和28.71%，低于文献报道的34.70%[10]。这说明一定比例无效干扰选项存在是不可避免的客观事实，遵从现有的严格的试题开发程序，可以维持无效干扰选项比例在同行业的相当水平。

4.2 实测数据的提示对发现非技术性因素产生的无效干扰选项更有意义

虽然命审题过程和无效干扰选项修改过程的原则是相同的，但是在实测数据的提示下，专家有能力发现无效干扰选项不足的原因。本研究中无效干扰选项修改原因中少数（6题，18.75%）为选项本身说法不科学、过于绝对或长度不同等技术性因素，多数（26题，81.25%）为选项内容干扰性不强。其中技术性原因可以通过不断强化技术培训和命审题要求尽可能避免，但知识内容方面的干扰性不强原则上不属于试题缺陷，在审题过程中不容易被发现，因此实测数据的提示作用非常重要。

4.3 针对性修改可减少无效干扰选项比例

选出的32道题无效干扰选项比例为63.28%（按照修正标准36.72%），修改一个最差选项后无效干扰选项比例下降为50.00%（按照修正标准30.47%）。

被修改干扰选项的选择率中位数从0.009提高到修改后的0.044，说明针对性对无效干扰选项进行修改，可以明显增加无效干扰选项的选择率，减小无效干扰选项比例。

4.4 减少无效干扰选项比例可提高试题质量

试题的区分度是衡量试题质量的一个重要指标。本研究中试题区分度从0.215提高到0.259，说明通过修改无效干扰选项可以显著提高试题区分度，改善试题质量。同时研究发现，相对于鉴别指数，点二列相关计算得出的区分度修改前后变化差异没有统计学意义。这说明，相比中间分数段人群，高分人群和低分人群对干扰选项修改所带来的难度变化更加敏感，也就是说，通过这种方式修改后的试题，能更好地区分开高分考生和低分考生。

4.5 减少无效干扰选项比例会增加试题难度

本研究中32道试题修改前试题难度为0.676，修改后为0.637，试题难度增加，有效选项数量增加会导致试题难度增加，这与相关研究结论一致[13]。但是修改前和修改后试题难度呈现高度正相关，说明试题难度虽然受有效干扰选项数量的影响，但更取决于试题题干、正确选项和考核能力内容等其他因素。另一方面也说明即使考核的知识点和表述方式不变，干扰选项的干扰性也可以影响试题难度，也可以成为调节试题难度的途径之一，也可以为题库建设和管理提供新的思路。同时，通过反馈实测数据对无效干扰选项修改的过程，也必将帮助专家减少命题过程中无效干扰选项的产生，提高通过控制选项来控制试题难度的意识，提升试题开发能力。

本研究的不足在于仅对每道试题中的一个选择率最低的无效干扰选项进行了修改，降低了专家修改难度，也可能对后续研究结果造成一定影响，如果对全部无效干扰选项进行修改，结果可能有所不同。但本研究仍然认为，通过分析实测数据来识别并修改无效干扰选项的做法是可行的，即使只选择性地修改其中一个选项，也可以提高试题的质量。

利益冲突　所有作者声明无利益冲突

作者贡献声明　李国建：组织实施、撰写论文；卢燕：统计学处理；何惧：提出论文构思、总体把关；刘茂伟：总体把关、审订论文；冯攀：审订论文

参考文献

[1] Gierl MJ, Bulut O, Guo Q, et al. Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review [J]. Rev Educ Res, 2017, 87(6): 1082-1116. DOI: 10.3102/0034654317726529.

[2] 李萍，苏小兵. 地理单项选择题干扰项功能分析：基于测试结果的评价[J]. 地理教学，2021(14)：16-20. DOI: 10.3969/j.issn.1000-078X.2021.14.006.

[3] 王来民. 追寻英语阅读理解干扰项踪迹，破解选项迷津[J]. 教学考试，2022(39)：26-29.

[4] 刘泽军，赵明德. 识别“雷区”巧用排除法：基于高考选择题的思考[J]. 教学考试，2022(44)：13-15.

[5] 刘拓，张佳慧，辛涛. 多项选择题中干扰项信息的利用[J]. 心理学探新，2015，35(3)：261-265.

[6] Rogausch A, Hofer R, Krebs R. Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: a simulation and survey [J]. BMC Med Educ, 2010, 10: 85. DOI: 10.1186/1472-6920-10-85.

[7] Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules [J]. Appl Meas Educ, 1989, 2(1): 37-50. DOI: 10.1207/s15324818ame0201_3.

[8] 冯攀，何佳. 医师资格考试单项选择题最优干扰选项数量分析[J]. 中国考试，2019(3)：48-53. DOI: 10.19360/j.cnki.11-3303/g4.2019.03.008.

[9] Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis [J]. BMC Med Educ, 2009(9): 40. DOI: 10.1186/1472-6920-9-40.

[10] 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 [J]. Adv Health Sci Educ Theory Pract, 2019, 24(1): 141-150. DOI: 10.1007/s10459-018-9855-9.

[11] Kilgour JM, Tayyaba S. An investigation into the optimal number of distractors in single-best answer exams [J]. Adv Health Sci Educ Theory Pract, 2016, 21(3): 571-585. DOI: 10.1007/s10459-015-9652-7.

[12] Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? [J]. Educ Psychol Meas, 1993, 53(4): 999-1010. DOI: 10.1177/0013164493053004013.

[13] Abdulghani H, Ahmad F, Aldrees A, et al. The relationship between non-functioning distractors and item difficulty of multiple choice questions: a descriptive analysis [J]. J Health Spec, 2014, 2(4): 148. DOI: 10.4103/1658-600x.142784.

（收稿日期：2023-05-09）

（本文编辑：张学颖）