# 教学质量和教学效果的评价方法

## —多选题考试法简介

上海第一医学院医学教育研究室编

卫生部科教司统考办公室

一九八三年三月

# 前　言

这本册子是我室受卫生部科教司的委托，为适应1983年医学院医学专业毕业生业务统考，分区举办多选法考试讲习班的需要而编写的。本册所采用的资料主要参考Hubbard.J.P 编著的“Measuring Medical Education”一书和Guilber.J.J 编著的“Educational Handbook for Health Personnel”一书，以及刘秉勋、梅人朗和刘立民三同志在卫生部科教司主办的1982年多选法命题工作讨论会上所作的专题报告讲稿。为了向各医学院校提供可供参考的典型试题，我们将卫生部科教司1982年部属高等医学院校医学专业毕业生试行业务统一考试的考题200道作为附录刊出。由于我们水平有限、取材上的缺点和错误在所难免，请读者提出批评指正。

**上海第一医学院医学教育研究室**

# 目　录

# 第一节　教学质量和教学效果的评价方法

## ——医学考试总论

### 一、考试评价的目的

考试和评价的目的是为了阐明受教育者在教学过程中是否达到了预定的教育目标，评定教学质量和各种教学方法的效果，更加具体地说，评定教学质量和教学方法的效果是为了：

1．为学校提供反馈，为改革教育，改进教学管理提供依据。

2．为教师提供反馈，有利于教师判明教学质量和各种教学方法的效果。检查教学内容是否合适；哪些问题学生已掌握了，哪些还未掌握。

3．为学生提供反馈，学生可以通过考试来评价个人学习上的成就和存在问题，哪些问题还不了解。

4．为决定升留级或是否授予学位提供客观的依据。

5．对社会负责，考试作为测量教育产品的一种尺度，保证向社会和公众提供合格的医生。

6．考试与评价手段也常常作为进行科学研究的一种方法，有助于教育工作者阐明医学教育的规律性。

### 二、考试评价制度的发展

古代的医学考试制度早在我国的汉朝就开始形成，公元前3年，汉元帝刘奭就建立了考校从官的制度，共分四科，医师就是从官之一，当时从民间通过考校医官的标准是质朴、敦厚、逊让、有行。到了宋朝，医学考试制度日益完善，1102年开始把医学教育划归国子监（相当于教育部）管辖。当时的太医署对医师实行分科培养，对学生的学习成绩考核采用三舍法，即分成上舍、内舍和外舍三级，成绩合格者升级，不合格者留级，有缺陷者经

过补习进入下一级，同时还规定习医的时间最高为七年，不能超过九年，据说这是世界上首先在医学教育中实行的制度化的医学考试。

欧洲的医学教育开始于九世纪，而医学院校的发展主要还是文艺复兴以后。十二世纪以后，牛津、剑桥、巴黎等大学兴起，这时开始了学位，因而开始建立考试，由于当时受宗教势力影响很大，所以考试内容除内外科外，更多地偏重于考哲学、修辞、占星术和教义。

欧洲文艺复兴时期以后的医学教育虽然随着大学的兴起，教学上开始重视自然科学的教学，但相当大程度上仍以“带徒培训”的方式来培养医生。因此，当时对医学生的考核实际上是通过师傅鉴定，由从师写个证明就算合格。如英国第一所医学院建于1123年，学生就是以学徒身份在医疗实践中培训，学制为七年，结业后由老师授予在伦敦地区行医的证书。

近半个世纪以来，由于医学科学技术的发展，医学知识不断地积累，要求学生必须学习的知识也越来越多，所以，如何评定学生的知识和技能已成为教育家们的重要研究课题。与此同时，由于教育科学的发展，教育目标分类学的出现，也为考试方法的设计提供了理论依据。鉴于这两方面的影响，医学考试不仅在内容上，而且在方法上也进行了重大的改革，特别是自七十年代开始，由于电子计算机被用于考试以后，考试方法和教育测量技术更向现代化方向发展。许多同志至今很难理解古代科举式的考试，也许很多同志也不了解最新测量技术，教学上仍习惯于采用问答法、填充法，是非法，即使有些学校采用了多选考试法，由于理论上实践上都没有对这些方法有比较深入的了解，往往导致了考试的失败。因此，在介绍医学考试技术前，有必要介绍一下考试的心理学问题。

## 三、考试设计的教育心理学基础

教学过程系由教育目标、教学计划和评价三部分所组成的互相影响互相联系的连续过程。因此，评价过程是根据教育目标来决定的。例如教学计划中必然会涉及到基本知识、基本技能和职业态度，那么，根据教育目标的要求，一种考试方法的提出主要应涉及哪一方面，这时一定要根据教育目标来决定，因此，我们有必要了解医学教育目标的分类学问题。

根据现代教育学理论，Bloom首先提出了教育目标的分类学问题，Bloom按学习的心理活动过程，把学习过程分成三个领域，即认识领域(cognitive domain)，精神运动领域(Psychomotor domain)和情感领域(affective domain)，这三个领域涉及哪些方面呢？根据WHO专家委员会讨论，它主要涉及：

1．认识领域（知识）

(1) 基本术语、各种论据、概念、原理、规律、方法和过程方面的知识；

(2) 对各种论据和概念的理解；

(3) 理解和解释各种资料与数据的能力；

(4) 解决各种问题的能力；

(5) 总体情况的分析；

(6) 综合的能力

2．精神运动领域（技能）

(1) 采集病史，向病人提出各种问题的技能；

(2) 进行体格检查的技能；

(3) 使用各种实验室和医疗器械的技能；

(4) 对患者进行系统观察和处理的技能。

3．情感领域（态度）

(1) 认识对病人应负的医疗责任；

(2) 关心和考虑到病人及其家属；

(3) 与同事以及卫生队伍其他人员进行有效合作的能力；

(4) 医疗上各种措施的应用及其局限性的认识；

(5) 对医疗上各种保护性措施的认识；

(6) 调查研究与续继教育的愿望。

根据学习目标分类学，Bloom把上述三个领域按照学习活动的心理过程分成不同的级别，以后McGuire又对Bloom的分级方法进行了简化，现介绍如下：

(一)认识领域

Ⅰ级 认识

Ⅱ级 理解

Ⅲ级 应用

Ⅳ级 分析

Ⅴ级 综合

Ⅵ级 评价

McGuire把Bloom分类法简化为：

Ⅰ级 回忆力和理解力

Ⅱ级 普遍化的能力

Ⅲ级 对一个熟悉问题的解答能力

(1) 对数据的解释 (2) 应用

Ⅳ级 对一个不熟悉问题的解答能力

(1) 数据的分析，(2) 特殊应用能力

Ⅴ级 评价能力

Ⅵ级 综合能力

(二)精神运动领域

Ⅰ级 模仿

Ⅱ级 操作

Ⅲ级　精密性

Ⅳ级　有机组合或共济运动（把一系列动作协调起来）

可以简化为：

Ⅰ级　模仿

Ⅱ级　控制

Ⅲ级　自然而然

三、情感领域

Ⅰ级　接受或注意

Ⅱ级　应答或反应

Ⅲ级　价值观化（行为显得坚持和恒定）

Ⅳ级　理论化（在自己头脑中建立了指导自己行为的价值观）

Ⅴ级　价值或价值体系的性格化

可简化成：

Ⅰ级　接受

Ⅱ级　反应

Ⅲ级　性格化

教育目标分类学在考试评价上的作用就是，要根据教育目标的要求来制定测量方法，一种理想的考试方法应能根据教育目标分类学的要求来选择测量的对象，确定测量的深度，也就是说，教育测量是在教育目标分类学的不同水平上进行。例如考查学生学习基本知识的情况时，根据学习计划，其考试的范围应属于什么领域，希望这次考试能测量到哪一级，是记忆、是理解、还是更高水平的分析综合能力。在课间考试中，往往是低水平的考试，主要了解学生对知识的记忆和理解，如果毕业考试，就要考更高水平的分析综合能力、评价能力。由此可见，由于智能活动可以被精细地区分成三个不同的领域，而每一个领域又可以分成

不同水平，故到目前为止，还没有一种测量方法能全面考核学生成绩的所有方面。

**四、制定考试方法的基本要求**

正如上面已经谈到了，根据教育目标分类学的原理来设计考试方案，这是对学生进行客观评价的基本出发点，由于不同方法所测量的范围不同，因此，在选择或制定任何一种考试方法时均应注意下列基本要求：

1．有效性

作实际应用的测量方法应当精密和有效，不应有外来因素的干扰，例如测量综合能力时，其他因素不应与综合能力相提并论，以免使其他能力的测量同综合能力的测量混淆起来。如去年的统考试题共分两部分，第一部分是基本知识的记忆和理解，第二部分是医疗问题的分析和基本知识的应用，如果题目出得好，就能有效地测量学生的知识和能力，如果出得不好，最后就不能反映学生的实际水平。

“方法的有效性”（“内容的有效性”）是指考试中所采用的方法能有效地用于测量预定的事物和行为，如果考试内容不能被所采用的方法进行有效地测量，则在评价结果时就会失去有效性。

“预期有效性”是指学生在一次考试中，某一方面所取得的成绩能否用来测量另一方面或另一情况成绩的可能性。例如根据某一临床学科的成绩来推测其基础科学和预防医学方面的成绩。去年的统考中，我们在设计命题方案时就考虑到题目的预期有效性，例如内外妇儿的考题，要求命题时要适当联系基础医学和预防医学，其目的就是考虑到通过一次考试能同时了解其他学科的知识水平，就是通过某一临床考题来了解学生基础科学与预防医学知识学得怎样？

2．可靠性

可靠性是指不同条件下对学生所测得的成绩是否一致，即某一考试方法所提供的结果的可重复性，例如用市尺、公尺和英尺所测得的长度，换算结果都一致的话，那么测量的结果认为是可靠的，在一次考试中，主考人和试题可以不同，但考试成绩应当相同，或者同一方法重复或分成几次进行均能取得同一水平。此外，大家知道，如果命题上存在许多漏洞，如暗示、猜测、冒险都会降低考试的可靠性。

可靠性是保证有效性的必要条件：成绩只有可靠才能有效，换一句说，有效必然是可靠的，但可靠的结果不一定是有效的，因此，不可靠的结果会影响有效性，其区别就在于可靠性是一个统计学上的概念，它可以用可靠性系数来表示，而有效性则是根据经验上来推断的，所以，可靠性就是考试成绩的可信程度。目前各院校普通反映存在成绩贬值或叫分数膨胀的现象，有个学校反映，学生最不重视的课程在考试中都得高分，这说明考试的可靠性很低，可信性很差，这种情况就要作进一步分析。

3．客观性

客观性是指每一种测量方法所取得的成绩，怎样才算“优良”或“及格”，即一次考试中各评卷人之间达到一致性的程度。

例如传统的问答题，有人曾作了研究，发现一份卷子由不同人批改，其成绩可以相差10分以上，如下表所示：

| 评分差异数 | 人数 |
|---|---|
| 0～1 | 1 |
| 2～3 | 6 |
| 4～5 | 20 |
| 6～7 | 34 |

| | |
|---|---|
| 8～9 | 10 |
| 10～11 | 3 |
| 12～13 | 2 |

为了消除评分人的主观因素，考试成绩必须由更多的人来评分，然后取其均数，以达到客观的要求。那么一份问答题要多少人批呢？有人研究结果是：

| | | | |
|---|---|---|---|
| 作文 | 78人 | 哲学论文 | 127人 |
| 数学 | 13人 | 物理 | 16人 |
| 拉丁语解释 | 19人 | 英语 | 28人 |

4．相关性

即选题标准与测量方法的一致性，其涵义与“内容有效性”几乎完全相同，相关性也是一个统计学上的概念，例如1982年考试成绩能否客观地反映实际水平，我们要计算各校学生在统考中所取得的成绩是否同学校的成绩相关。

5．平衡性

即同每一教育目标分类学相关连的题目数与教育目标测量方法理想分量的比例是否协调。

例如平时在考试中可以把重点放在低水平的认识领域，而毕业考试时就要考虑高水平认识领域的要求。

6．同一性

即考题与教学内容之间的协调程度，就我们来说，考题是否按教学大纲的基本要求来出，如果离开了教学大纲规定的内容，则考试结果必然缺乏一个可供测量的标准。

7．区别性

即一道试题在考试中能区别高分学生与低分学生的能力，用区别指数来表示。理想地说，测量方法的每一部分，在特殊情况下也应当能区别出学习优秀的学生与学习差的学生。当然，一道

题目100%的学生答对，似乎这道题目的区别性很差，实际上，没有区别性的考题也可反映两种情况，一是考题内容是教学上反复强调的，每一个学生必须掌握的，故这次考试结果说明全体考生达到了预期要求，另一种情况也可能是考题出得很坏，它不能区分优秀的和差的学生，甚至在评分时明显地有利于低分学生，所以区别性很差，严格说，这样的考试题在评分时应当除去。

8．效率

即单位时间内，用这种测量方法能保证学生完成最大考题数的特征。例如一道问答题往往20分钟才能答完，而一道多选题按美国的标准只要45秒，所以，采用多选题的效率大大提高，单位时间内效率的提高意味着考试范围的扩大，因此，评定学生实际能力的客观性也随之加大，现在问答题使用越来越少的原因之一是效率低，考核的深度虽较多选题大，但范围却很狭小。

9．题量

按照 Spearman-Brown 氏的论点，用增加与原测量方法所规定的相等数量的考题，则测量方法的可靠性几乎可以无限止的增加。所以，一次考试中的题量不能太少，有人说多选法考试范围很小，但它的优点是题量可以成倍增加，故它的可靠性是其他方法所不及的。

10．特殊性

即按照测量方法的特征，一个优秀的学生即使没有学过也能取得预期的成绩，也就是说考试方法本身应当有利于发现冒尖的学生。

**五、各种考试方法的优缺点比较**

所谓考试方法，就是指考核某一特定目标时所采用的应试材料和应试方法，目前，医学考试中所采用的方法，按照应答类型，总的说来可以分成两大类，即自由应答型(free-response type)和

固定应答型(fixed-response type)。所谓自由应答型考试，就是方法本身允许学生可以用自己的语言或行动来对某一问题作出回答，包括论述题考试、小论文、简单的直接回答题、口试、填充题或者由老师对学生的操作过程作出没有具体规定的评定。而固定应答型的应试材料则是事先计划好的，对考题的答案也是主考人事先一致商定的，这种方式包括由主考人用口试或在实际环境下进行的是非题、多选题、改进型书面问答题、病人处理问题(PMP)、标准核对表和评分标准等。

总的说来，自由应答型考试具有要求学生作出即时回答的优点，回答时学生要充分运用本人所掌握的知识。其缺点是：这种方法用机械评分是不可能的，并且常常会受到应试者言语流利与否，或主考人主观印象等外界因素的影响。

固定应答型的优点是可以用电子计算机进行评分，也可以由任何主考人按照标准答案来评改试卷，不受主观因素影响。这种方法也有缺点：学生可以用猜测的方法来回答，备选答案也可能提供暗示。无论怎样，由于这种考试方法不断得到改善，使它可以针对教育目标分类学中某一部分或某一水平来考查单一的或复合的能力，它可以考记忆的，也可以考记忆加上理解的，也可以考记忆、理解与应用，另外，这种方法也可以客观地和方便地评定成绩，所以，这种考试方法已得到了广泛的使用。

以下就两类考试中的几种方法加以比较：

| 考试方法 | 考核范围（按照Bloom教育目标分类学） | 优　　点 | 缺　　点 |
| --- | --- | --- | --- |
| 论述题 | 测验知识的回忆水平、理解（理解力）及应用（解决问题的能力） | (1) 可以看出学生如何选择适当知识来回答的能力；<br>(2) 除考核知识的回忆力外，还可看出更高级的解释力和解决问题的能力；<br>(3) 可以了解论述方法及表达力；<br>(4) 命题比较容易。 | (1) 因考题数有限，只能限于对考试部分进行了解；<br>(2) 评分容易产生主观性；<br>(3) 容易忽略学得多并不一定是好的解答这一正确观点；<br>(4) 文字通顺和出现错别字有可能影响成绩<br>(5) 平时了解"好"或"差"学生对同样答案不一定能得到同样评分（Holo效应）<br>(6) 评卷时间长，难以保持评分标准；<br>(7) 命题不佳，提问不清楚，学生难以回答，因而难以了解学生的思考力；<br>(8) 考题数少，对教学与学习的反馈作用较少。 |

| | | | |
|---|---|---|---|
| 口　试 | 测验认识领域回忆力以上的各种能力（解释、应用、分析、综合、评价） | (1) 可以了解学生解题的思考过程；<br>(2) 根据应试者回答的情况，可以面对面灵活地处理；<br>(3) 可以造成很调和的气氛；<br>(4) 可以由几名主考人同时评分；<br>(5) 可以使用X线照片、心电图、标本、模型、组织标本、录象和录音带等视听教具进行提问。 | (1) 评分标准难以保持一致和客观，考试的可信性较低；<br>(2) 各应试者的考题不同时，其可信性较差<br>(3) 化的时间较多，考生多时困难更大；<br>(4) 在应试者之间可能互相通报考题；<br>(5) 教师有可能作启发性的问话；应试者情绪比较紧张；<br>(6) 能言善辩的学生和沉着冷静的学生易被评为高分，沉默寡言和谨小慎微的学生则易被评为低分；<br>(7) 易出现Holo效应；<br>(8) 有的考题事先没有作好充分准备而是偶然提出的；<br>(9) 教师对口试大多也不习惯；<br>(10) 考题少，对师生的反馈也少。 |

| | | | |
|---|---|---|---|
| 客观考试（多选题、复合是非题、填充题） | 测验知识的回忆水平，较单纯的解释力及某种程度的解决问题能力也能测量 | (1) 单位时间内的题量可以非常多、从而保证了试题的广泛性；<br>(2) 可信性较高。考试的客观性对可信性很大，可信性又和考题数及考题质量有关，因此，不能说客观考试就是可信性高的考试；<br>(3) 评分时可以防止由于字迹好坏、错字、Holo效果等偏见；<br>(4) 评卷容易，也可用机械进行；<br>(5) 易于调整考题的难度<br>(6) 考题数多，好的试题可以重复使用；<br>(7) 试题分析比较容易；<br>(8) 考生多时非常经济；<br>(9) 师生双方都可得到详细反馈。 | (1) 命题要有专门的技巧，而且较费时；<br>(2) 只能测量认识的低水平；<br>(3) 任意猜测也有可能作出正确的回答，在5个备选答案上总会有20%的机率；<br>(4) 对很浅的印象，模糊不清的概念有可能引起记忆；<br>(5) 为了通过这种考试，学习方法有采用死记硬背的危险；<br>(6) 给予在实际上不可能存在的提示。在临床上要求自己独立判断的问题，在备选答案中有可能得到提示；<br>(7) 应试技术可能影响成绩，在选择正确答案的思考过程中，从某种程度上来看有一些诀巧。 |

| | | | |
|---|---|---|---|
| 模拟考试（创造模拟实际环境来测量解决问题的能力） | 综合性解决临床问题的能力（解释、应用、分析、综合、评价） | (1) 象多选题一样，非常客观，<br>(2) 编成程序以后，可以由计算机来控制，让学生通过一系列模拟实际环境的程序一步步地进行；<br>(3) 可以系统地观察学生综合评价的能力；<br>(4) 程序可以编成几天、几个月或从发病到死亡或康复的全过程；<br>(5) 师生可以即时得到反馈；<br>(6) 可以由演员或生过病的人来模拟；<br>(7) 可用作自学和自我检查的一种工具，对继续教育更加适用。 | (1) 制订考题相当化时间、相当复杂；<br>(2) 考卷要有专门的技术进行处理；<br>(3) 应试者和模拟者都要经专门的训练 |