

试卷质量评估与分析系统的构建与实现

倪海儿^{1,2}, 裘晓华³, 魏丹毅^{1,4}

(1.宁波大学 本科督导委员会, 浙江 宁波 315211; 2.宁波大学 海洋学院, 浙江 宁波 315211;
3.上海海洋大学 信息学院, 上海 201306; 4.宁波大学 材料科学与化学工程学院, 浙江 宁波 315211)

摘要: 根据教学测量学原理和统计分析方法, 构建试卷质量评价与分析的指标体系. 用灰色关联评价对试卷质量进行综合评价, 并在EXCEL VBA中构建了试卷质量分析系统, 实现试卷质量的分析与评价. 以期帮助改进教学方法, 提高教学质量, 提高教学效果, 以及提供试卷质量状况的有效信息服务.

关键词: 试卷质量; 评估指标; 灰色关联度; 综合评价

中图分类号: G644

文献标志码: A

文章编号: 1001-5132 (2016) 03-0118-05

试卷是实现考试功能的工具, 高质量的试卷能真实、全面地反映学生的学习情况, 实现考试的评定、区分、预测和激励导向等功能. 对试卷及其考试结果的分析与检验还可以获取教学效果的反馈信息, 发现教学过程中存在的问题, 分析诊断造成问题的原因, 为改进教学工作, 提高教学效果提供参考. 但在教学与管理过程中, 往往更注重试卷的评定作用, 而忽略了对试卷的分析与评估, 因此难以从中了解教学效果和教学反馈信息, 从而无法很好地发挥其对教学应有的指导作用. 因此, 客观、全面地评价试卷的质量, 充分挖掘考试所反馈的教学信息, 可为检验教学效果、改进教学提供科学手段和必要依据. 本文根据教学测量学原理和统计分析方法, 构建试卷质量评价与分析的指标体系, 并在EXCEL VBA中构建了试卷质量分析评价系统, 实现试卷质量的评价与分析, 拟为教师提供自我分析与诊断的工具, 为教学管理部门提供教学管理与质量控制的依据.

1 评价指标的确定

从测验目标来看, 课程考核属于成就测验, 是在课程结束后, 对教学目标所预期的教学效果做概括性的测定, 对学生的学科能力给出评价. 因此

试卷质量关乎两个最基本的问题: 一是试卷是否反映教学大纲的要求, 二是试卷能否准确反映学生的实际水平. 通过对试卷质量的评估和考试成绩的统计分析可反映测验是否达到考核的目标及其教学的效果.

1.1 试卷的质量指标

对一个测验而言, 测验结果的可信性、有效性是评价其质量的重要指标, 同时要实现考试的区分、评定功能, 难度、区分度和覆盖度也是质量评价的基本要素, 因此本系统选用信度、效度、难度、区分度和覆盖度作为试卷质量评估的基本指标^[1].

试卷作为测量学生学科能力的工具, 测量结果的可信性程度是衡量试卷质量的最基本要素. 由于测量中存在各种随机因素影响, 任何测量总存在随机误差, 导致测量结果的随机波动, 这种随机误差愈大, 测量的稳定性愈差, 测量结果的可信性程度就愈差. 而信度是测量结果可信性程度的指标, 其通过对测量随机误差的衡量及结果的稳定性程度来反映可信性程度. 方差分析可根据观察值的变异原因, 将观察值的变差进行分解, 从而给出随机误差的估计^[2]. 对于一组测验分数, 用总离差平方和来衡量数据的总变异, 记为 Q_T , 测验分数的变异是由被试个体间的变异、题目之间的变

异和它们两者交互作用造成的, 因此根据方差分析的思想, 可将 Q_T 分解为被试个体之间、试题之间以及它们两者的交互作用的离差平方和, 即:

$$Q_T = Q_p + Q_s + Q_{p \times s}, \quad (1)$$

式中, Q_p 为被试个体间的离差平方和; Q_s 为试题间的离差平方和; $Q_{p \times s}$ 是它们两者交互作用的离差平方和, 可作为测量误差的估计. 一个好的测验应最大限度地将被试变差归结为个体之间的差异, 即测量误差尽可能小. 荷伊特用误差均方与个体间均方的比来衡量信度^[1]:

$$C_{xx} = 1 - \frac{M_{p \times i}}{M_p}, \quad (2)$$

式中, C_{xx} 为信度指标; $M_p = Q_s / f_s$ 和 $M_{p \times i} = Q_{p \times s} / f_{p \times s}$ 分别为个体的均方和个体与题之间交互作用的均方, 其中 f_s 和 $f_{p \times s}$ 分别是 Q_s 和 $Q_{p \times s}$ 的自由度. 信度指标是否有效, 可通过对题目间和个体间效应的 F 检验来确定^[2], 如果个体间的效应显著 ($p < 0.05$), 则可认为这个信度标准可靠. 当信度标准可靠时, C_{xx} 的最大值为 1, 则测量误差为 0.

效度是衡量一个测验能达到其测量目的有效程度的指标. 本系统采用内容效度和效标关联效度来衡量. 内容效度指考试内容与教学大纲要求达到的目标的一致性程度, 通常由专家和教师评判. 考虑到这种评判带有较大的主观性, 本系统采用试卷的覆盖面来反映^[3]. 效标效度是指测验用于测定个体特定成效的有效性, 而要判断这种测定的有效性, 就必须找到独立于测验结果的、客观的标准, 即效度的标准-效标, 并以此为标准, 来衡量某一测量的有效程度. 但在实际应用中, 效标往往难以给出. 在课程考核中, 效标效度是用来反映该测验能反映学生真正掌握和应用该学科知识的能力. 本系统选用学生平时成绩作为测量的效标. 为衡量测验结果与效标的关联程度, 通常采用相关法及显著差异法等^[4]. 由于课程考核成绩一般采用百分数, 所以本系统采用 Pearson 相关系数来衡量测验分数与效标测量的关联程度. Pearson 相关系数 (r_{xy}) 定义为:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (3)$$

式中, n 为被测个体数; x_i 和 y_i ($i = 1, 2, \dots, n$) 分别为被测个体的测验分数和效标分. r_{xy} 取值在 $(-1, 1)$ 之间, r_{xy} 值愈大, 测验结果与效标测量的一致性程度愈高, 测验的效度愈高. 同时可以通过对 r_{xy} 进行显著性检验^[2], 测验效度是否有显著意义.

难度是测定被试完成测验的困难程度. 适当的难度才能达到测验的评定和区分作用. 因被试的得分高低反映了试题的难易程度, 因此难度系数 (E) 可由测验平均得分占总分的比例计算^[4]:

$$E = \frac{\bar{x}}{100}, \quad (4)$$

式中, $\bar{x}$ 为测验的平均分. 各题的难度系数 (e_i) 为:

$$e_i = \frac{\bar{x}_i}{x_{i(\max)}}, i = 1, 2, \dots, p, \quad (5)$$

式中, $\bar{x}_i$ 为所有被测在第 i 题上的平均得分; $x_{i(\max)}$ 为该题的满分值; p 为题目数. 以上定义中 E 值愈大, 试卷的难度愈小. 为与通常的观念一致, 本系统中难度系数用 $1 - E$ 表示.

区分度是测验区分或鉴别被试者在该学科能力水平上差异程度的度量. 为便于使用, 采用高低分数法^[1]. 对测验分序列按从高到低排序, 计算前 27% 人 (高分组) 的平均分和后 27% 人 (低分组) 的平均分, 测验的区分度 (D) 为 2 个平均分之差在满分中所占的比例:

$$D = \frac{\bar{X}_H - \bar{X}_L}{100}, \quad (6)$$

式中, $\bar{X}_H$ 和 $\bar{X}_L$ 分别为高分组和低分组的平均分. 各题的区分度 (d_i) 可由下式计算:

$$d_i = \frac{\bar{X}_{i(H)} - \bar{X}_{i(L)}}{F_i}, \quad (7)$$

式中, $\bar{X}_{i(H)}$ 和 $\bar{X}_{i(L)}$ 分别为第 i 题前 27% 人和后 27% 人的平均分, F_i 为该题的满分值.

覆盖度表示试题对教学大纲规定内容的覆盖程度. 各题得分的相关系数可反映各题间内容的相似程度, 如果各题间相关系数大, 说明各题内容相关性大, 否则说明各题的内容相对独立, 因此试题的覆盖度可通过试卷中各题间的相关系数矩阵来反映. 第 i 题和第 j 题的相关系数为:

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ik} - \bar{x})(x_{jk} - \bar{x})}{\sqrt{\sum_{k=1}^n (x_{ik} - \bar{x})^2 \sum_{k=1}^n (x_{jk} - \bar{x})^2}}, i, j = 1, 2, \dots, p, \quad (8)$$

式中, x_{ik} 和 x_{jk} 分别为第 k 个个体在第 i 题和 j 题上的得分, 相关系数的平均值为:

$$\bar{r} = \frac{p(p-1)}{2} \sum_{i=1}^{p-1} \sum_{j=i+1}^p r_{ij}. \quad (9)$$

试卷的覆盖度为 $1 - \bar{r}$, 覆盖度愈大, 试卷的内容效度愈高, 它的最大值为 1.

1.2 试卷质量的统计分析

信度、效度、难度、区分度和覆盖度是从教育测量学角度评价试卷的质量. 测验结果的统计分析一方面可从测验分数的分布来反映教学效果, 另一方面也可以反映试卷的质量. 本系统对测验结果进行描述性统计分析, 包括测验结果的平均水平、离散程度和分布形态. 平均水平用测验分的均值表示, 它能反映教学效果整体水平的高低或试卷的难易程度. 通过对均值检验可以推断教学是否达到了预定的教学目标, 也可以比较不同被测单元(如班级)在学科能力上的差异或不同教学方式之间的差异. 离散程度包括标准差、标准误差、最大值、最小值、极差, 反映了学习效果的差异的程度. 由于学生的知识基础、学习能力、努力程度等有差异, 学习效果会有差异. 因此一定大小的标准差是教学正常的反映, 但如果标准差过小, 且均值过高或过低, 则反映试卷设计中难易程度掌握不好, 区分度不够; 反之若标准差过大, 说明两极分化大, 反映教学环节中存在问题. 由于人的心理特性受众多因素影响, 因此在教学正常的情况下, 反映在测验上的成绩应呈正态分布. 偏度 S (Skewness) 和峰度 K (Kurtosis) 能反映成绩分布的大体情况, 它们可分别由以下两式计算:

$$S = \frac{(\sum_{i=1}^n (x_i - \bar{x})^3) / n}{(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2})^3}, \quad (10)$$

$$K = \frac{(\sum_{i=1}^n (x_i - \bar{x})^4) / n}{(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2})^4}. \quad (11)$$

式(10)及式(11)中, x_i 为第 i 个个体的测验分. 当测验成绩服从正态分布时, $S \sim N(0, 6/n)$, $K \sim N(3, 24/n)$, 因此可用它们的观察值来考察测验分数偏离正态性的程度. 进一步地将 1 个被测单位

(如班级)作为来自某种教学方式下总体的 1 个样本, 可通过这个样本来推断其总体是否为正态分布, 从而提供教学过程、试卷质量状况的信息. 为保证检验结果的可靠性, 本系统采用峰度和偏度的 χ^2 和 K-S (Kolmogorov-Smirnov) 2 种检验方法对测验结果进行正态性检验.

2 综合评价方法

灰色关联分析是通过序列曲线几何形状的相似程度来反映它们的关联程度^[5], 灰色关联度愈大, 两序列的变化过程愈一致. 因此可通过度量测验与目标值的关联程度来评价测验的优劣. 以试卷质量的基本评价指标(信度、效度、难度、区分度、覆盖面)和反映成绩分布形态的基本指标(峰度、偏度)为评价指标, 根据这些指标的意义选出它们的最佳值作为评价测验好坏的目标向量, 记为 $X_0 = (x_0(1), x_0(2), \dots, x_0(7))$, 记 $X_i = (x_i(1), x_i(2), \dots, x_i(7))$ ($i = 1, 2, \dots, m$) 为测验分数序列. 通过计算测验序列对目标序列的关联度对测验进行评价. 首先对 X_i 由下式计算它们的初值像 x_i' ,

$$x_i'(k) = \frac{x_i(k)}{x_i(1)}, k = 1, 2, \dots, 7, x_i(1) \neq 0, \\ i = 0, 1, \dots, m, \quad (12)$$

然后, 计算 X_i 与 X_0 对第 k 个指标的关联系数:

$$\gamma(x_0(k), x_i(k)) = (\min_i \min_k |x_0'(k) - x_i'(k)| + \\ \xi \max_i \max_k |x_0'(k) - x_i'(k)|) / (|x_0'(k) - x_i'(k)| + \\ \xi \max_i \max_k |x_0'(k) - x_i'(k)|), \quad (13)$$

其中, $\xi \in (0, 1)$ (本系统中取 0.5) 称为分辨系数. X_0 与 X_i 的灰色关联度则为:

$$\gamma(X_0, X_i) = \frac{1}{7} \sum_{k=1}^7 \gamma(x_0(k), x_i(k)). \quad (14)$$

可见 γ 取值在 0~1 之间, γ 值愈大, 说明该测验愈接近于目标值, 则该测验愈优. 因此将 γ 作为试卷质量的综合评价指标, 把评价结果分为 4 个等级, $\gamma > 0.75$ 为优, $0.5 < \gamma \leq 0.75$ 为良, $0.25 < \gamma \leq 0.5$ 为合格, $\gamma \leq 0.25$ 为不合格.

3 系统的功能与结构

成绩的录入与基本统计分析是系统的基本功能. 在输入考试人数和考题数量后, 系统生成成绩表结构, 然后进入成绩录入界面. 用户可在成绩录

入界面录(导)入或复制各题成绩、平时成绩, 系统自动生成基本统计结果, 包括总成绩以及各题的平均分、标准差、标准误差、最高分、最低分、极差、峰度、偏度等。

成绩的统计分析模块提供总成绩的频数统计、正态性检验, 并用图、表来表示成绩的分布, 同时计算难度、效度、信度、区分度以及相关系数矩阵和覆盖度。

试卷质量的综合评价模块用灰色关联分析给出试卷质量的综合评分及其评分等级。

打印模块生成、打印试卷质量分析报告。

为了保证成绩录入系统的安全性, 避免用户误操作引起的工作表和窗口的删除和修改, 对工作簿结构和窗口进行了保护, 以禁止用户添加或删除工作表, 或显示隐藏的工作表, 禁止用户更改工作表窗口的大小或位置。

分析系统的主界面如图1所示。其中, “使用说明”模块除了系统的功能说明及操作提示外, 还给出了各指标在试卷评价中的意义与评价标准, 便于用户对试卷中的各题进行更细致的评价与诊断。试卷分析报告的实例如图2所示。

4 讨论

效度是所测结果与欲测目标的一致性程度, 是评价一个测验优劣最为重要的指标。效度的验证有内容效度、结构效度和效标关联效度。因为结构效度所涵盖的范围包括内容效度和效标关联效度^[1], 因此本系统中采用内容效度和效标关联效度来测量测验的有效程度。效标关联效度是测验结果与反映被试真实水平或能力指标(-效标)一致性的程度。效标关联效度的评价中, 效标的选择最为关键, 不同的效标会产生不同的效度, 从而得出不同的效度评价。所以应根据测验的目的和性质来选择最佳效标, 使其能客观、准确地评价测验的效度。但在实际应用中效标往往难以获得, 因本系统是针对于课程考核进行的评价, 课程考核的目的是测验被试掌握和应用该学科知识的能力, 这样的能力应该体现在整个学习过程中。因此本系统选用学生的平时成绩作为测量的效标。当然这就要求在平时的教学过程中, 应全面、准确地记录被试的学习成绩。



图1 试卷质量分析系统主界面

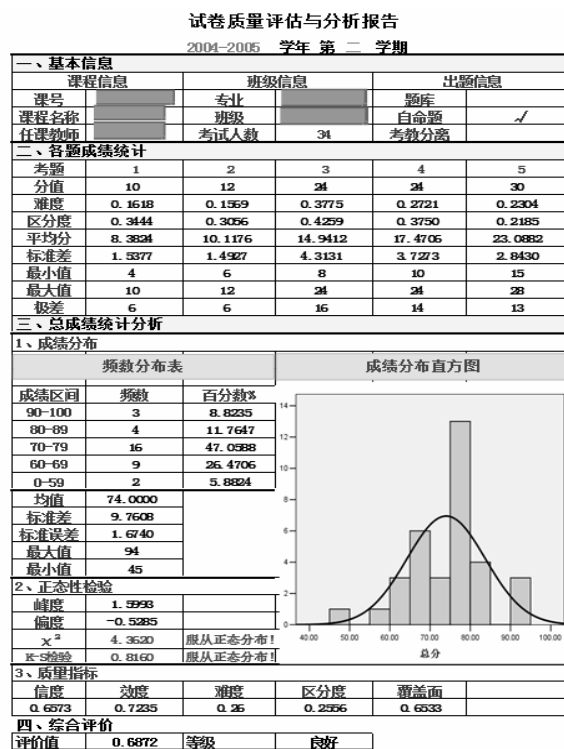


图2 试卷质量分析评估报告实例

为便于自我评价与诊断, 本系统给出了各个评价指标的参考标准。但标准的确定与测验的目的和性质有关, 因此参照一些学者提出的标准^[4,6-9], 并结合课程考核的要求给出。如考虑到课程测验要既能检验学生达到教学目标的程度又能体现学生能力的差异, 难度系数 $1-E$ 值不宜太大, 但为了区分与激励, 又不能太小, 因此把难度系数在0.25~0.35之间定为优。

灰色关联分析是基于序列间的相似性和接近性, 来评判不同序列间联系的紧密程度。它对样本容量和总体分布都没有要求, 尤其适用于小样本情况。因此本系统把灰色关联分析引入到试卷质量评价中, 通过测量测验与目标向量的关联度来

评价测验. 因目标值是一种理想状态, 离理想状态愈接近, 试卷的质量愈高. 因此目标向量的确定中, 信度、效度、区分度、覆盖度均取最大值, 难度系数取 0.3, 根据峰度与效度的分布特征, 峰度取 3, 偏度值取 0. 本系统中各指标是等权重的, 在进一步研究中, 也可以根据测验目的的不同对各评价指标赋予不同的权重. 在评价等级的确定中, 也可根据测验的不同要求, 确定相应的评价等级.

参考文献:

- [1] 黄光扬. 教育测量与评价[M]. 上海: 华东师范大学出版社, 2007:46-87.
- [2] 倪海儿, 钱国英. 概率论与生物统计[M]. 2 版. 杭州: 浙江大学出版社, 2013:121-164.
- [3] 杨继泰. 综合评价试卷质量[J]. 药学教育, 1997, 13(4): 36-38.
- [4] 张敏强. 教育测量学[M]. 北京: 人民教育出版社, 2005:80-134.
- [5] 刘思峰, 党耀国. 灰色系统理论及其应用[M]. 北京: 科学出版社, 2010:62-85.
- [6] 张雅波, 李媛媛. 试卷质量的分析评价模型[J]. 职业教育研究, 2008(5):51-52.
- [7] 刘新平, 刘存侠. 教育统计与测评导论[M]. 北京: 科学出版社, 2003:235.
- [8] 余仙菊, 彭如宽, 卢凤娟. 教育测量理论在高校课程考试中的应用研究[J]. 中国高等医学教育, 2006(8):23-25.
- [9] 詹向红, 杨雪, 芦锰, 等. 试卷分析是题库建设的重要环节[J]. 中国医药导报, 2008, 5(3):100-104.

Construction and Implementation of Quality Analysis and Evaluation System for Examination Papers

NI Hai-er^{1,2}, QIU Xiao-hua³, WEI Dan-yi^{1,4}

(1.Undergraduate Steering Committee, Ningbo University, Ningbo 315211, China;

2.School of Marine Sciences, Ningbo University, Ningbo 315211, China;

3.College of Information Technology, Shanghai Ocean University, Shanghai 201306, China;

4.Faculty of Materials Science and Chemical Engineering, Ningbo University, Ningbo 315211, China)

Abstract: By adopting the principles of educational measurement and the statistical methods, an index system is set for the quality evaluation and analysis of examination papers, whereby the grey relational analysis is used for the comprehensive evaluation of the paper's quality. By taking advantage of the Excel VBA feature, the quality analysis system is implemented and is put to test. The developed system is expected to provide the information about the teaching effectiveness and problem quality given in examination so as to improve teaching quality and enhance teaching effectiveness.

Key words: examination paper quality; evaluating indicators; grey relative relationship; comprehensive evaluation

(责任编辑 章践立)