试卷质量评价与分析

白瑞琳
（沈阳师范大学　辽宁　沈阳　110034）

【摘　要】评价试卷的质量对做好考试工作有着重要的意义，试卷质量的高低对反映教师教学水平和学生对知识技能的掌握程度有很大影响。试卷质量评价不仅可以推动试卷命题设计的改进，助力考试评价质量的提升，还能更好更准确检测出学习效果。本文从试卷整体的信度、效度、知识点覆盖率等进行评价，还从题目的难度、区分度进行评价。并对此次评价分析结果做出总结和思考。

【关键词】试卷；质量；评价

引言

检验教学质量最重要的手段就是成绩考核，而成绩考核能否真实反映教学质量的关键问题在于考核的命题（也就是考卷)的质量。[1] 对考试试题和试卷的质量进行分析。为此，本文对A省S市扬帆中学初二下学期期中试卷进行质量评价，并做出，有助于了解学生的学习情况，推动命题设计的改进，为教学提供反馈信息，及时发现教学安排和试卷中存在的问题，以便改进教学、提高试题质量，从而确保考试职能的有效发挥总结和思考。

1 对象及方法

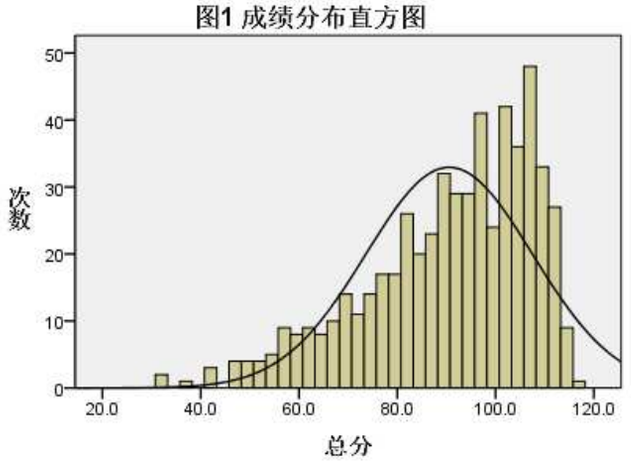
1.1 评价对象

2021 年 4 月，A 省 S 市扬帆中学初二学生期中考试试卷共 560 份，学生成绩情况如下图图 1。本次考试试卷共 86 题，满分 120 分，其中客观性单项选择题为 60 道，共 75 分；主观性简答题 20 道，共 25 分，书面表达 6 道，共 20 分。

1.2 评价方法

用 EXCEL 建立数据库，将各题的编号、得分、答对情况和每个学生的英语总成绩等项目录入计算机，以 SPSS.22.0 统计软件包进行数据处理和统计分析，对此次试卷进行定量评价。此外，参考教学大纲和教师总结知识点，利用文本分析法对试卷进行定性评价。

图1 成绩分布直方图



2 试卷整体评价

2.1 试卷信度

信度指的是测量结果的稳定性或可靠性的程度，亦即测量的结果是否真实、客观地反映了考生的实际水平。[2]

表 1 可靠性统计资料

Cronbach 的 Alpha	基于标准化项目的 Cronbach 的 Alpha	项目个数
.910	.963	3

同质性信度（通过克龙巴赫 Alpha 系数表示），若信度系数在 0.9 以上，说明量表的信度很好；若信度系数在 0.7-0.9 之间，说明量表的信度可以接受；若量表的信度系数在 0.5-0.7 之间，说明试卷信度一般，有些试题需要更换或改进；若信度系数在 0.5 以下，说明信度有问题，考试基本无效。根据表 2 可看出克隆巴赫系数为 0.910，信度系数在 0.9 以上，说明此次试卷信度很好。此外，为了保证试卷信度，我们还通过询问阅卷老师评分标准得出评分者信度，客观题由计算机机读卡评阅，主观题由每位阅卷老师使用同一标准，每张试卷由三位教师评阅。在评阅过程中不出现学生信息，且分差过大会进行再一轮评阅，很大程度上保证了此次试卷的信度。

2.2 试卷效度

效度是一个测试准确性和有效性的数量指标。效度的估计有多种方法，常分为三大类：内容效度、效标关联效度和结构效度。效标关联效度是测量试卷有效程度的主要方式，可以通过选择合适的效标，计算学生本次考试成绩与所选效标之间的相关系数，能够较准确地测试出学生掌握和运用所学知识的真实度。

表 2 相关

		期中成绩	平均成绩
期中成绩	皮尔逊（Pearson）相关	1	.983**
	显著性（双尾）		.000
	N	560	560
平均成绩	皮尔逊（Pearson）相关	.983**	1
	显著性（双尾）	.000	
	N	560	560

**. 相关性在 0.01 层显著（双尾）。

我们这里用皮尔逊系数表示效度，将前两次月考的平均成绩作为效标，通常认为相关系数在 0.4-0.8 较为理想。通常情况下通过以下取值范围判断变量的相关强度：相关系数在 0.8-1.0 表示极强相关，在 0.6-0.8 表示强相关，0.4-0.6 表示中等相关，而在 0.2-0.4 表示弱相关，至于 0-0.2 则显示出极弱相关或不相关。根据表 2 可看出，相关系数为 0.983，表示极强相关，所以此次试卷有效。

2.3 知识点覆盖率

知识点覆盖率指标是用以衡量试题基本知识点和类型题

覆盖程度的一项指标。U 表示知识点覆盖率指标；I 表示试题中所涵盖的基本知识点；R 表示按照学科教学大纲规定所要求的基本知识点。

这里结合教学大纲要求和教师总结知识点作为参考，期中考试范围为 1-4 单元，共有 71 个需要掌握单词和 90 个知识点，试卷中呈现的知识点共 105 个，我们用 U=I/R 计算得出 U=0.65。知识点覆盖率由学科组讨论，通常情况下，知识点覆盖率 0<U<0.7 时，认为试题不合格；0.7 ≤ U ＜1 时，认为试题合格。此次试卷知识点覆盖率小于 0.7，我们认为试题不合格。造成试题覆盖率没有达到合格的原因在于听力和阅读部分，听力部分题目中一些是比较简单的基础知识，不在此次范围内。而阅读部分有一半考查了课外阅读，虽然没有涉及到课内知识点，但是可以增加学生词汇量，帮助学生培养对于英语的敏感性，有助于学生了解不用国家的文化习俗，提高学生的语言综合运用能力，拓宽学习视野。

2.4 题目数量

试卷质量不仅和题目质量有关，也和数量有关。根据经验，一般控制方法是：预计中等程度的学生用 70％左右的时间可以完成。在经过询问学科老师后，初二英语老师表示此次试卷的题量和前两次月考基本一致，学生可以在规定时间完成，只有少数学生会完不成书面表达题。在这里我们再参考另外一种方法，在限定的考试时间内应保证有 90％的学生答完所有试题，所有学生都能答完全部试题的 90％以上。根据试卷成绩统计可以看出，客观题平均空白率为 0.07％，主观题中有 54 名学没有答完主观题部分。综合来说，此次试卷题目数量比较合格。

2.5 题型比例

对试题类型划分，是希望找到一种测量只能活动水平的合理试卷结构，从而提高命题的科学性、稳定性和可靠性。本次考试客观题全部为单项选择题，主观题为主观简答题和书面表达题。题型比例如表 3

表 3 试卷题型分析

题型	题量（％）	分值（分）
单项选择题	60（70％）	75
主观简答题	20（23.3％）	25
书面表达题	6（7％）	20

主客观题都有其优缺点，单一用一种题型的考试有一定的局限性，要全面测量考生的知识水平，应考虑二者的比例问题。根据张世俊的说法，大规模考试的客观题比例都要大一些，一般都不低于 50％，而平时的考试主观题的比例可以稍大一些。研究表明客观题主观题两者在 7：3-3：7 之间对学生成绩的影响无差异。此次考试客观题达到了 70％，主观题 30％，对学生的成绩影响无差异，但是此次考试为平时的期中考试，建议可以适当增大主观题的题量和分值。

3 试卷题目评价

3.1 题目难度

难度 P：指试题或试卷的难易程度，它是衡量试卷质量的一个重要指标参数。客观题：P=R/N，其中 P 表示试题的难度，R 为答对该题的人数，N 表示参加测验的人数。主观题用公式 P= /k 表示，为该题的平均分，k 为该题的满分。我们将客观性单项选择题按照顺序标为 k1-1，k1-2，k1-3... 主观性简答题按照题型标号为 z26-30,z66-70，z71-80，z81-85,z86。由于客观性单项选择小题数较多，且听力部分结果良好，这里不再呈现听力部分，只呈现笔试部分部分，笔试部分分为单项选择、完形填空和阅读理解三部分，完形填空和阅读理解部分较多，在这里只呈现单项选择，如下图表 4-1，主观性简答题如下图表 5。

表 4-1 单项选择难度区分度

题目标号	难度	区分度
k1-31	0.81	0.39
k1-32	0.726	0.46
k1-33	0.723	0.39
k1-34	0.819	0.15
k1-35	0.736	0.43
k1-36	0.785	0.56
k1-37	0.941	0.4
k1-38	0.734	0.39
k1-39	0.601	0.5
k1-40	0.831	0.48

非首次考试，可以用同学科前面的考试的统计难度作参照来进行预测。所以我们用前面考试难度作参考，从历次考试分析看，难度控制在 0.7-0.9 之间比较合适，有利于测量学生的真实水平，对及格率也有一定的控制。P ＜ 0.5 的试题过难，失分严重，应分析原因。我们可以看出 k1-37 难度在 0.941，过于简单，这道题目考的是 if 从句的固定使用，说明学生对这个知识点掌握的比较好，在后期学习中适当进行巩固即可。k1-39 题有一定难度，但并不过度。此题考察的四个短语的辨析，且题目中也出现了短语，学生在答题时可能不理解题意或混淆选项意思，容易答错。后期学习中应注意短语的意思并准确识记。

完形填空题目中 k1-42、k1-43、k1-44、k1-45、k1-46、k1-47、k1-48、k1-49 题的难度分别是：0.898、0.732、0.879、0.766、0.739、0.873、0.883、0.898，这八道题难度和区分度都处于正常系数中。但是 k1-41 和 k1-50 题过难，失分严重，特别是 k1-50 题，有很大的难度，这道题目是完形填空中的最后一道题，四个选项分别 Happily,Suddenly,Luckily,Clearly 四个副词，虽然四个词的中文语义学生都能识记，但是这个选项在最后一段承上启下的位置，也是一句话的开头，容易造成上下文不理解，曲解文章本意，造成选错选项，建议教师在今后的教学中加大此类题目的练习力度。

阅读理解共 15 道题目，其中 10 道题都处于正常水平，5 道题难度不在正常范围内。k1-51，k1-52，k1-58，k1-59,k1-60 这五题相对来说比较简单，k1-63 难度在 0.5 以下，难度相对比较大，这道题目考察的是指示代词的用法，需要学生完全理解文章意思，且四个备选项在文中都有所涉及，也不易使用排除法。

其余的 k1-53、k1-54、k1-55、k1-56、k1-57 难度为 0.859、0.883、0.692、0.769、0.844；k1-61、k1-62、k1-64、k1-65 难度为 0.745、0.717、0.714、0.686 都处于正常平均的状态。

表 5 主观性简答题难度区分度

题目标号	难度	区分度
z26-30	0.38	0.46
z66-70	0.88	0.25
z71-80	0.73	0.39
z81-85	0.79	0.44
z86	0.57	0.34

主观性简答题中 z26-30 题难度为 0.38，在 0.5 以下。题目过难，失分过多，z16-30 题满分为 5 分，平均分为 1.91 分。这几道题为听力填空题，不仅要听出来空缺单词，而且要把单词填写正确，一定程度上相对听力单项选择题增加了难度，建议在听力练习中多练习此类题目，而且要保证单词的正确记忆，这样才能保障在听出空缺单词的基础上把单词写正确。其他主观题题目较为良好，z86 为书面表达题，可以在固定句式和书写上适当得分。

3.2 题目区分度

区分度指测验对考生实际水平的区分程度，用 D 表示，区分度高的试题（ 或试卷），能较好地鉴别考生的实际水平，使得实际水平高的考生得高分，实际水平低的考生得低分。区分度是评价试题（ 或试卷） 质量、筛选试题的主要指标和依据。此次试卷的区分度如上图表 4-1，表 4-2，表 4-3 和表 5。美国测验专家伊贝多尔（L. Ebel）于 1965 年提出 D 值的评价标准为：D ≥ 0.40 试题优秀；D ＝ 0.30-0.39 试题良好；D ＝ 0.20-0.29 试题尚可，需进行修改；D ≤ 0.19 试题较差，应摒弃或进行修改。但是此次评价的是期中考试试卷，期中考试的目的主要是为了检查，不是筛选，区分度的意义不大。由区分度表可看出，k1-34 区分度为 0.15，应适当进行修改。此题考察的是 afford 的固定用法，根据句意选择 v-ing 形式或者是不定式形式，有 82％的学生选择了不定式形式，由此可见，学生虽然知道 afford 的固定用法，但是却没有完全掌握，在应用方面还远远不够。k1-50 题区分度在 0.07，62％的学生选择了 C 选择项 Luckily，14％的学生能理解文意，区分 Luckily 和 Clearly 哪个单词用在文章中更合适。此题目一定程度上区分了学生的阅读文章和理解上下文逻辑的能力，但是可以进行一定的修改，在文章文意上更明确一点。

单项选择中其他题 k1-31、k1-32、k1-33、k1-35、k1-36、k1-37、k1-38、k1-39、k1-40 区分度分别为 0.39、0.46、0.39、0.43、0.56、0.4、0.39、0.5、0.48；完形填空中 k1-41、k1-42、k1-43、k1-44、k1-45、k1-46、k1-47、k1-48、k1-49 区分度为 0.5、0.41、0.55、0.38、0.64、0.5、0.45、0.43、0.51；阅读理解中 k1-51、k1-52、k1-53、k1-54、k1-55、k1-56、k1-57、k1-58、k1-59、k1-60、k1-61、k1-62、k1-63、k1-64、k1-65 区分度为 0.27、0.42、0.45、0.46、0.45、0.54、0.43、0.34、0.39、0.42、0.23、0.6、0.29、0.41、0.32。这些题都属于合格题，区分度都在 0.19 以上，不仅可以检验学生学习，也可区分不同水平的学生。

3.2.1 评分标准

试卷的评判依据提前制定的统一评卷标准， 客观单选题由计算机读卡评阅， 主观题按照公平公正原则， 按照流水作业程序， 采取集体封闭方式进行评阅。这一方面主要看书面表达题评分标准，书面表达题评分标准共有 6 点，分别是：

(1) 整篇作文满分 15 分，其中内容 6 分，语言 6 分，结构 3 分。

(2) 内容贴切，句子流畅，用语准确，加整体印象分 1 分。

(3) 不满 70 个词，少 1-5 个词扣 0.5 分，6-10 个词扣 1 分。

(4) 所有给出题纲涉及的内容，每少一项扣 3 分。

(5) 每个拼写、大小写、标点符号错误扣 0.5 分；同一个错误扣分总和不超过 2 分。

(6) 语法错误每项扣 1 分，同一错误扣分总和不超过 2 分。

根据此次试卷书面表达题评分标准可看出，虽然评分标准在内容、字数、拼写和书写方面做了一定的细则，但是没有明确内容要点。同时，也没有分出文章各档次给分范围和要求，不够详细，在这里如果能增加一档文、二档文评分标准和内容要求会更完善。

3.2.2 思考及意义

3.2.2.1 推动试卷命题设计的改进，助力考试评价质量的提升

试卷质量评价的正向作用，对试卷设计工作有着促进作用。即使是有经验的教师，也可能会设计出无效的试题，但在设计中可能没有注意到。试卷题目的难度分析教师可以知道 哪部分对于学生来说过难，哪部分过易这样他们可以设计出更为实用的试卷。除此之外由于语言类考题的特殊性，有经验的教师还可以根据学生在试卷中易犯的错误设计出改错题。教师在今后试题设计中节省大量的时间，而且也会帮助他们确定教学重，提高教学质量。从而间接促进教学的改进，发挥考试评价对教学的导向作用。

3.2.2.2 检测自身阶段学习的效果，调整学习复习计划的执行

期中考试是学生找准学习方向的依据，能促进学生在学习上取得一定的进步。期中考试是给学生回顾、反思自己的机会，否则很多学生会形成“期末突击”的应试方式。学生通过期中考试这种外力驱动，不得不进行阶段性复习和总结，在试卷质量好的情况下能更准确地检测前半学期学习的效果。并且通过小结，学生根据自身实际情况可以判断自己学习方法是否合理，发现自己学习过程中的问题所在，以便在下学期有针对性的进行调整。

参考文献：

[1] 杨萍，郭卫华．关于试卷质量评价的进一步讨论 [J]. 信阳师范学院学报 (自然科学版),1997(03):102-106.

[2] 胡中锋，李方．教育测量与评价 [M]. 广州：广东高等教育出版社 ,2000:31-58.

作者简介：

白瑞琳（1997.5—），女，汉族，籍贯：河北邯郸人，沈阳师范大学教育科学学院，20 级在读研究生，硕士学位，专业：课程与教学论，研究方向：课程论，教学论。