第27卷 第4期
2005年7月

湘潭师范学院学报(社会科学版)
Journal of Xiangtan Normal University(Social Science Edition)

Vol. 27 No.4
Jul. 2005

语言测试成绩分析与命题质量

尚巾斌
(湖南科技大学 外国语学院,湖南 湘潭 411201)

摘 要:尝试运用测试学的有关理论和统计手段,对湖南湘潭工学院99级非英语专业使用新教材的192位本科学生第三学期大学英语课的期末考试成绩进行抽样分析,阐述如何通过对考试成绩进行量化分析来评估试题的命题质量及对教学的意义。

关键词:测试;测试成绩;信度;区分度;难度;评估
中图分类号:H003 **文献标识码**:A **文章编号**:1009-4482(2005)04-0167-03

测试是英语语言教学中再熟悉不过的一个教学环节。对于教师,这种熟悉往往演变成对测试的麻木。多数情况下,教师阅卷评分后将学生的成绩上报即万事大吉。熟不知,考试本身只不过是一种评估手段,我们希望通过考试来了解或获得有关信息,而这些信息的首要来源是考试成绩。因此,考试的结束并不意味着教学环节的终结,教师还需要对测试成绩进行分析以获取量化的信息反馈,这对进一步完善教学,提高命题质量都大有裨益。

一 测试成绩分析的手段

测试成绩分析可以根据需要在多个层次上进行。比如,我们可以进行统计获得所有参考人或某一年级、班级总成绩及各组成部分的集中量,即均数、中数、重数和频数,以了解成绩分布的集中趋势,或者说学生掌握知识或某一知识技能的总体情况。此外,我们也可以进行差异量(如全距、标准差等)的统计来了解个层次上学生成绩分布的离散状况。我们还可以统计考试成绩的正太分布,直观地了解学生成绩是否合理。一般说来,集中量、差异量和成绩的正太分布的统计只能让我们获得对测试本身及学生知识掌握情况的概括认识。这些也是我们通常采用测试成绩分析手段。更重要的是,通过测试成绩的分析教师可以获得题目的难易度值、区分度和整个试卷的信度指数,从而让我们对该测试的试题质量有一个科学的认定,有利于完善教学,提高命题质量。

这里,本文拟将通过对湘潭工学院99级非英语专业本科生第三学期大学英语课程的期末考试成绩进行抽样分析,具体说明测试成绩分析之于命题质量评估的意义所在。

二 考试说明

A.目的与性质:本次测试是湘潭工学院99级非英语专业本科生使用新教材的6个自然班级第三学期的大学英语课其末考试。参加本次测试共192人。本次测试的目的是为了检查学生在完成了三个学期的大学英语课程后,是否达到了《全国大学英语教学大纲》的相关要求。从这个意义上说,本次考试属于尺度参照性考试,因为本次考试以教学大纲的标准作为测量的依据,以大学英语教学内容作为测试的基础。但本次考试又有别与一般的成绩考试。在内容上它不局限于以某教材的内容为准绳,而是参照《全国大学英语教学大纲》的要求。从这个意义上讲,它兼有水平测试的特征。严格地说,它属于尺度参照性水平测试的范畴。

B.试题型与试卷结构:本次测试采用与目前全国大学英语四级笔卷考试类似的形式,主观题与客观题相结合。试卷总分为100分,由听力理解、阅读理解、词汇与结构、完型填空、写作等五部分组成。各部分所占比例分别为20%、40%、15%、10%、15%。其中客观题占整个试卷总分的85%;主观部分,即作文占15%,要求学生在规定的30分钟时间内完成一篇不少于100个单词的命题作文(段首句给出),整个测试时间为120分钟。

二 考试信度与试题质量

1.考试信度

考试信度是评价与保证考试质量的重要指标。对于任何一种有效考试来说,考试信度是必不可少的。那么,什么是信度?简单的说,考试信度是指考分的一致性和可靠性。考试如果要成为有效的测量手段,信度是很关键的。Bachman认为对信度的研究主要是解决这样一个问题,即"考生成绩受旨在测试的语言能力以外的因素影响的程度,并将这些因素对分数的影响降低到最低程度"。要使考分有效的前提是考分必须是可信的。因为,任何一个考试最终是通过分数来表述其结果的,如果分数不可信,我们就无法说

收稿日期:2005-01-30
作者简介:尚巾斌(1970-),女,土家族,湖南吉首人,湖南科技大学外国语学院英语一系教师,上海外国语大学英语语言文学专业研究生,主要从事语言学、应用语言学研究。

考试有效地评估了学生的语言技能。

2.测试结果与信度评估

测试学界一般通行三种测定信度的方法:重复测试法,平行卷测试法和对半分析法。重复测试法更似容易,但在实际操作中有一定的难度。首先,施行重复测试法要控制两次考试相隔一定的时间。间隔时间太短,学生则可能凭记答题;太长,学生的语言能力则可能发生变化。其二,控制相隔时间段内的教学内容,以保证学生的进展方面的一致性。因为,这种方法的可靠性基于一个同步性的假设上,既在两次考试之间学生在学习上都没有或都获得新的进展。如果,在此期间一部分学生的学习成绩发生变化,而另一部分则保持原状,那么,第二次考试的分数与第一次相比会出现差异,因而,两组分数之间就缺乏稳定性。在这种情况下,用重复测试法定信度,得出的考试信度数据会偏低。

平行卷测试法是让学生做两套试卷,随后分析考试的结果。通过两组分数的比较而求得考试信度;分数组之间的一致性将决定考试的信度。两份试卷可一次先后完成,或隔天完成。平行卷测试法对 A—B 卷的制作要求较高.由于这两份卷子被视为完全等同的试卷,因此在考试内容、题目难度及其坡度、试卷长度、试题数量、绝考时间、甚至于题目顺序方面都必须吻合,这给教师命题任务提出了挑战。

以上两种测定信度的方法,在实际操作中的可行性上都有一定的难度.比如时间间隔问题和试卷的完全一致性问题.为了避开这两个难题,语言测试界通常采用对半分析法来确定考试信度。对半分析法是把一份试卷作为两个相对独立且相应的部分,通过这两个部分分数的比较以获得整个试卷的信度。这两个部分分数的一致性越高,试卷的信度也就相应地越高。对半分析法也称作一种求内部一致性的方法(internal consistency)。语言测试领域有多种求内部一致性的方法,这里我的采用 Spearman Brown 的修正公式(correction formula,有时也称作 Spearman Brown prophecy formula)。(数据参见附表)

通过对半分析法相关算法,我们得出这样的结果:

(1)客观题部分的信度系数为 0.749。

(2)整个试卷的信度系数为 0.651。

根据测试学有关理论,信度系数分节范围为 0~1 之间。Lado 认为如果是客观题,信度系数在 0.9 以上为佳。从我们抽样检测的结果来看,本次测试整个试题客观部分的信度比值为 0.676。应该说这个结果不太理想。测试学理论要求整个试卷的信度系数至少应在 0.7 以上为较合理,本次考试整个试卷的信度系数与要求略有差距。那么,问题出在哪里呢?考试的信度会受到各种因素的牵制。理论上说,信度的建立首先是题目的质量问题。不论是主观考试还是客观测试,题目的质量的优劣直接关系到考试的信度。其次,题目的难度也会影响到信度。在此,我们进一步采取定量分析方法,通过对本次考试试卷内各项目之间相关的分析,对题目难易度、区分度作抽样检测来评估试题的质量。我们从参加本次考试的 8 个自然班级中,抽取两个自然班级的期末考试成绩(以下简称 A,B 组)作为分析样本。

三 测试成绩组成的各部分之间的相关分析与试题质量

1.相关分析

在这部分的分析中,我们分别对构成两组学生成绩的五个部分分数的频数分布、集中量、差异量及正太分布情况进行逐步统计,以期从中找出各部分之间的相关性。(统计结果见附表)为试题的质量及学生对知识的掌握情况提供量化的指征。

从统计结果中,我们发现:

1)A、B 两组成绩在听力理解、阅读理解、完形填空、写作四部分表现出来的趋势与总成绩趋势一致,即 A 组在这四部分的成绩离散度明显大于 B 组。但在词汇结构题部分,A、B 两组成绩的差异分数几乎相等,(A 组:0.28 B 组:0.27),即 A、B 两组成绩在这一部分的离散度几乎没有差距。这一趋势与其余四部分的趋势相反。

2)A、B 两组成绩分别在写作部分的分数均过于集中;两组中分别 90%的分数集中在 8~10 分。这种分布显然不太合理。难道是学生的写作水平惊人地一致?抑或是试题本身的难易度、区分度不合理?或是判分中有失客观公正?要找出原因对于这两部分,我们需要对于这两部分作个案分析。

2.问题分析

A.词汇结构部分

对于导致 A、B 两组成绩在这一部分试卷上反映出的反常趋势,笔者有两种假设:

1)两组学生的词汇、语法能力本身无差别;

2)词汇结构部分题的难易度、区分度不合理;

a 难度与区分度分析

根据测试学理论,难易度值在 0~1 之间,数值越小说明题目越难,反之则表明题目容易。一般来说,题目容易会导致考试的整体难度降低,考试就无法拉开考生之间的距离;题目太难会使考试的整体难度上长,同样无法区分不同水平的考生。比较理想的作法是把题目的难易度控制在 0.3 和 0.7 之间。同时,要合理安排难易度比例,并使题目的难易度呈坡度形状,即把容易的题目放在前面,把难的题目放在后面。我们不妨先对这部分题的难易度、区分度进行统计。该部分共 30 道多项选择题,我们抽取所有的奇数题共 15 道(总数的一半)作为检验样本。

统计结果显示,15 道题中有 9 道题的难度系数在 0.3 与 0.7 之间,占 60%;5 道题的难度系数在 0.7 以上,占 33.3%;只有 1 道题的难度系数为 0.17,占 6.7%,也就是说,93.3%的题在难易适中和偏容易的范围内,只有 6.7%的题属难度大。这说明,整个这部分试题偏容易。另外,试题的难易排列无序。

再看看它们区分度。测试学中规定区分度值在 0~1 之间,1 表示完美的区分度;0 表示有效区分度。怎样的区分度指数属于可接受范围呢?在一般性况下,命题人员可以接受 0.4 或高于 0.4 的区分度指数对于低于 0.4 的题目则要慎重对待,进行修改或撤换。从上表的结果看,15 道题中有 10 道题的区分度指数在 0.4 以下,占 66.7%;其中有 6 道题的区分度在 0.2 以下,占 40%;15 道题中 5 道题的区分度在 0.4 以上,但最大的区分度指数也只有 0.54。这些数据说明,词汇与结构部分 2/3(66.7%)的试题的区分度在可接受的范围之外;其中,占总数 1/3(40%)的试题的区分度很低(区分度指数在 0.2 以下)。这一结果表明,学生的相应的这一方面的语言水平因为试题的区分度太低而没有被区分开。

b 各案分析

获得题目的难度和区分度指数后,我们可以发现那些题目有问题(如偏难、区分度低等)。对于这些题目,我们需要分析每个题目选择项的被选情况,以便深入了解学生的

答题过程,从而进一步探索语言习得的过程或语言能力的性质、构成因素等。对于那些有问题的题目,选择项分析可以让我们找到问题的症结所在,提供修改或撤换题目的依据。以本次测试的47题为例,47题的难度指数为0.17;区分度为0.18。数据表明该题太难,所以区分度小。那么,该题到底难在那里?学生的问题在那里呢?让我们从难度指数的统计中寻找答案。

47. The family——in southern France.

A) reside B) inhabit C)inhabitant D)leave

该题正确答案为A)reside。从统计中我们发现,高分组22人中除6人答对该题外,其余16人均选了B)inhabit,无人选C)或D),而低分组22人中无一人答对该题,其中13人选B),9人选C),无人选D)。这些统计结果至少说明两点:(1)高分组和低分组中无一人选择D选项,该选项形同虚效,没有真正起到干扰作用,需要修改,(2)绝大部分学生选B)inhabit而无人选A)reside,这表明学生对reside一词的用法没有掌握,或对reside与inhabit两词的区别没有掌握,教师应着重讲解reside的用法及其与inhabit的区别。

B 写作部分

统计中我们发现两组学生的写作成绩过于集中,分布明显欠合理。从抽样检测中,我们发现问题如下:

首先,本次考试的写作部分为短文写作,其性质和要求与论述题较为接近。这种类型的作文题一般应由三个部分组成:提示(prompt)、题目(topic)和要求(reqirement.)。本次作文部分只有题目(topic),没有提示(prompt)和要求(reqirement.)。笔者以为,让学生知道评分要求有利于他们了解作文的要求,可以针对性地进行写作。

另外,抽样分析显示,写作成绩过于集中,问题主要出现在评分方面。对于档次相差很明显的作文,在评分上没有得出体现。原因可能是多方面的。一则,可能是出提者没有制定明确细致的评分标准。评卷者采用的是印象评分法,由于缺乏细致的评分标准,评卷者容易出现评分前后不一致的现象;二则,评分效度不合理。在正式阅卷开始前,先要组织阅卷人员从试卷中挑选出能代表评分档次或个分数段的样卷(benchmar scripts)。在改卷过程中同意认识、同意理解。再者,为保证评分效度,阅卷时一般应采用两人互阅制,最后分数为两人分数的折中。这样可在一定程度上缩小阅卷中的分数偏差。而抽样分析显示,一份试卷只有一人判分,而且送分现象普遍而明显。这也许是因为评分者总是希望更多同学能够及格的缘故。

C 结论

从以上信度分析、试题各项之间的相关分析以及具体的个案分析结果来看,本次测试的信度系数没有达到最佳度,也就是说考试分数解释学生语言能力的功能不是很理想。抽样分析中所列举的多项的评估数据表明,信度欠佳与部分试题的质量及难易度、区分度有直接关系;个案分析中表明词汇与语法结构部分试题命题欠合理,难易度区分度掌握欠妥当,个别试题需要修正。另外,试题写作部分试题质量和评阅均有问题,但主要是评阅欠客观。此外,在抽样检测的两个平行班中,A组学生成绩离散度相对大一些,说明水平参差不齐,而B组学生水平整齐一些。因此,出卷者应根据反馈的信息进一步提高试题的质量,以提高整个考试的质量,有利执教的教师因材施教。

四 启示

当今语言测试界已达成这样一个共识,即考试与教学有密切了联系,并具有后效作用。如何考与考什么会对日常的教学与学生的学习产生一定的影响。本文运用测试学的一般原理,通过实例重点分析说明了测试结果的评估对于考试信度和试题质量的检测的意义。以上在测试学理论指导下的测试结果分析给我们这样一些启示:

(1)测试的结果不能忽视。对测试结果进行科学的评估是教学中的一个必要的环节。这个环节能提供关于测试本身以及被测试的对象的科学认识。因此,测试结束后,命题人员和执教教师要对测试成绩在多个层次上作科学的分析与评估,并在此基础上总结反馈信息,以利于今后教学。

(2)测试命题中考试规范的编写不能忽视。考试大纲不能代替考试规范。考试规范是单个或一系列有关的正式文件。对考什么和如何考作出详细的规定。就对象来说,考试规范的使用者是考试命题人员或考试评估人员,而考纲是面向广大的考生或教师。就内容而言,考试规范详尽地表述考试所考的部分、所考试的语言范畴、具体语言技能、一定评分标准的使用等。而考纲只概括性地介绍考试的有关方面。或许有人会说,考生人数少的考试无需考试规范一类的正式文件,教师可以自行确定考试内容和项目等。但我们认为,即使是小规范考试,如年级或班级的期中、期末考试,即使考试的涉及而不广泛,教师仍需写出考试规范。考试规范的制订是考试效度的一个必要保证。考试规范对考试内容也应有详细,明确的表述,这有助于命题教师在命题过程中做出内容更有代表性或针对性。考试规范对评分标准的明确规定也有助于评分的一致性。本次测试中存在的一些问题,与缺乏考试规范有直接的关系。

参考文献:

[1] 邹 申.英语语言测试[M].上海:上海外语教育出版社,1998.

[1] 邹 申.简明英语测试教程[M].北京:高等教育出版社,2000.

[2] 尚巾斌.英语语言测试中的效度评估及其意义[A].语言与文化论文集[C].上海:上海外语教育出版社,2001.

[4] 大学英语教学大纲[M].北京:高等教育出版社,1999.

[4] Hughes, Arthur. **Testing for Language Teachers**[M]. Cambridge University Press,1989.

[5] Bachman. **Fundamental Considerations in Language Testing**[M].上海:上海外语教育出版社,1997.

[7] Bachman, Lyle and Palmer, Adrian S. **Language Testing in practice**[M]. 上海外语教育出版社,1997.

[8] Lado, Robert. **Language Testing**[M]. Longman GroupLimited,1961.

[9] Osferlind, sfenen, J. **Constructing Test Items**[M]. Kluwer Academic publisher,1989.

[10] **Oller, John W Jr. Language Tests at School**[M]. Longman Group limited,1979.

(责任编校 朱正余)