文章编号:2095-0365(2017)02-0068-04

# 计算机辅助普通话水平测试“命题说话”项评分误差控制

杨会永，　张　宁

(石家庄铁道大学 人文学院 ,河北 石家庄　050043)

摘　要:国家普通话水平测试“命题说话”项主观性较强,评测误差不容易控制。设计开发一套计算机软件,对评测员的测评工作进行实时监控,力图将由主观性导致的测评误差控制在最小程度,监控主要从测试员试评误差控制、测试员与测试员集体之间一致性对比误差控制、测试员本人一致性误差控制、不同评测点之间误差控制、管理员抽查监控误差控制等六个方面进行。这套程序将对目前正在使用的评测系统产生积极的辅助作用。

关键词:普通话; 命题说话;水平测试; 误差控制

中图分类号:G434　文献标志码:A　DOI:10.13319/j.cnki.sjztddxxbskb.2017.02.13

目前国内计算机辅助普通话水平测试所使用的是科大讯飞公司研制开发的“国家普通话水平智能测试系统”,该系统基于国家普通话水平测试大纲,可较准确地对考生“命题说话”之外的三个题型进行自动评测。“目前的语音评判技术还无法对考生在没有文字凭借的情况下的第 4 题说话项目上的普通话水平进行评判”[1]。这样,第 4 题“命题说话”则由该系统录音后由测试员基于该系统网络平台在网上进行评分,这较之以前的人工集中评测有很多的方便之处。但须承认的是,由于“命题说话”项的评分受评测员自身水平的影响较大,具有较强的主观性,故对考生的评分存在程度不等的误差。所以,如何有效减小“命题说话”项的评分误差,真正测出考生实际的普通话水平,切实提高普通话水平测试的公信度,已成为计算机辅助普通话水平测试要解决的重要问题之一。我们结合自身多年的普通话测试经验,并学习借鉴国内一些水平考试的有效做法,对计算机辅助普通话水平测试“命题说话”项的评分误差控制做了研究。

我们对目前“命题说话”项的评分误差控制主要是研究开发了一套计算机辅助普通话水平测试“命题说话”项人工评分误差控制系统,该系统能对评测员的评测结果进行实时监控,能将由于评测员主观性导致的评测误差控制在评测过程之中。

## 一、测试员试评误差控制

“命题说话”项评分标准中,“语音标准程度”较其他五项其主观性最强,该项评分共分为六档,评测员的评分误差多在这项中出现。为使评测员对六档评分标准有比较准确的把握,我们设计从往年“命题说话”项语音数据中每档选取 50 人的说话数据(评分确定比较准确的),通过排列组合编排 100 组试评题目,每组 5 个说话语料,每组题目要照顾到“语音标准程度”的不同档次,即一组题目尽可能显示多类型语音面貌。如果时间允许,可以多编排几个说话语料,因为“当其样本容量增大时,测量的概化系数 $E\rho^2$ 和可靠性指数 $\varphi$ (类似于 CTT 中的信度)都会提高”。[2] 测试员每

收稿日期:2016-12-20

作者简介:杨会永(1967— ),男,副教授,博士,研究方向:汉语言文字学。

本文信息:杨会永,张　宁. 计算机辅助普通话水平测试“命题说话” 项评分误差控制[J]. 石家庄铁道大学学报:社会科学版,2017,11(2):68-71.

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

天上网评测前均要为一组题目评分、对比后才能评测,以帮助测试员尽快找到语音面貌归类感觉。

## 二、测试员与测试员集体之间一致性对比误差控制

这项功能是监测某测试员的评分与测试点全体测试员评分分布是否一致,如评测员集体评分分布为正态分布,而某测试员与正态分布有误差,则表明该测试员评测与集体评测存在误差。对此类误差,有多种方法进行控制,我们依靠评分曲线来控制。假如一个测试点有考生 3 000 人,测试员 10 名,考生测试之后测试员开始评测,每天所有测试员的评测分数会自动生成一个曲线,每个测试员的评测分数也会生成一个曲线,就会看到单个测试员评测分数曲线和评测员集体评测分数曲线的比较图,随着评测人数的增多,就会越来越清楚地看到每位评测员的打分倾向,是与集体打分分数分布曲线吻合还是或高或低。上述单个评测员的评分数据,与评测员集体数据相比较,如在误差允许范围内,可视为合格。否则,监测系统将在该评测员的屏幕和监测员的屏幕上出现提示警告信息,以提醒评测员及时修正。这个方案我们是参考了一些地区高考作文评分电脑监控的方法提出的,陈佳民指出:“有了评分常模,即对评分标准理解方面多了一个数量模式作为参照,就大大提高了对评分标准的理解的准确性和一致性。”[3] 我们这套程序所提供的测试员集体的分数曲线就是一个评分常模,可以作为个体评测员的评分参考。

图 1 是我们所研制计算机辅助普通话水平测试“命题说话”项人工评分误差控制系统的一个截图。该图是“语音标准程度”档测试员集体评分曲线和刘荣贵测试员评分曲线的比较,从图 1 中可以看出,刘荣贵测试员的得分分布基本在集体分布的左侧,表明其控制标准偏严格,考生得分偏低。“词汇与语法规范程度”等五项也可用同样的方法进行检测,不再详述。

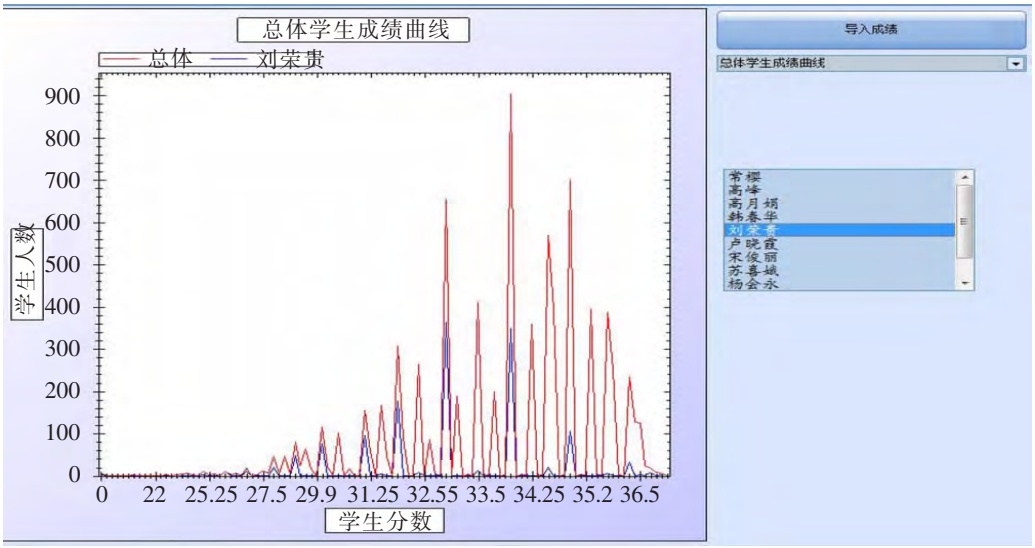


图 1　普通话测试评测员个体打分与评测集体打分对比图

## 三、测试员本人一致性误差控制

这项功能的目的是检查测试员在评测时掌握标准是否宽严适度,始终如一,是否随意性大。控制方法是监测员可定时随机抽取某评测员已经评测完成题目的百分之几来由其本人复评。比如某评测员今天评测的 3 个说话语料,过两天仍安排他复评。如同一说话语料同一测评员不同时间段的评分误差超过一定限度,则可说明该测评员评测标准掌握不准,需提醒或取消其评测资格。该方法在评测之初可多使用,包括复评比例也可提高,待评测稳定后复评比例和次数可相对减少。

## 四、不同测试员两评、三评误差控制

“在普通话测试中一个主要的可能误差,就是评分者之间或评分者本身内部(例如:对甲、乙部评分松紧不一)的不一致性”。[4] 为了避免这种人为的误差,同一说话语料往往由两个评测员评分,一般取其平均分作为得分。如果两个评测员评分相差较大时,这时就会产生一个阈值,“阈值是指两

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

个评分之差的绝对值达到原先定义评测误差的最大允许值"。[5]两评相差的绝对值超过某一值时,评测系统会自动分配给第三位评测员评测。评测的阈值可由评测管理部门根据实际情况确定。这项功能目前科大讯飞的测试系统也有,但根据我们研制的控制系统检测,科大讯飞平台这项功能尚不完善,两个评测员的评测分数有些超过某一阈值的并不能检测出来,在这方面还需完善。

## 五、不同评测点之间误差控制

目前普通话测试中"命题说话"项的评测是由各单位自己负责评分的,如河北的各高校的评测都是每个学校的评测员评测本校的测试者。一个单位的测试员一般十几个,这些测试员绝大多数都是兼职测试工作的,往往是每年集中评测一至二次,这就很可能造成一个单位的评测员评测水平参差不齐,也就很有可能出现一个单位总体评测分数偏高或偏低的情况。为避免这种情况发生,保证一个地区评测的公正性,我们在系统中设计了不同评测点之间的误差控制,即在评测时可参考同地区不同测试点的评测分数。如同在石家庄的省属重点高校的学生就可以作为评分互为参考的对象。这种误差控制主要由省级普通话测试管理中心负责,可以随机抽取某个评测点的已评说话语料由另一个评测点的评测员再评,如误差率在允许的范围内可视为合格,如误差率超过一定的阈值则需提醒被抽测评测点注意评测标准。

## 六、管理员抽查监控误差控制

本方法主要由省级评测管理中心和各测试站管理人员操作,主要使用管理、监控手段从整体上来减少误差。如可采用简单随机抽查某位测试员的评测成绩;还可使用分阶段抽样方法,如可抽取第一天和某一天的评测成绩做比较,看前后标准把握是否一致;还可自动抽样检测,即检测系统可以根据统计结果自动抽取某位评测员的评测成绩同整体评测成绩相比,以发现评测中存在的问题;还可使用等距抽样方法检测,即间隔抽样,如可抽取考号尾数是 2 的考生成绩进行复评。

## 七、其他功能

### (一)分级率统计功能

我们这套程序可以将一个单位的所有考生的普通话测试等级进行统计,以图表形式直观展示一个单位考生的普通话测试等级,见图 2。

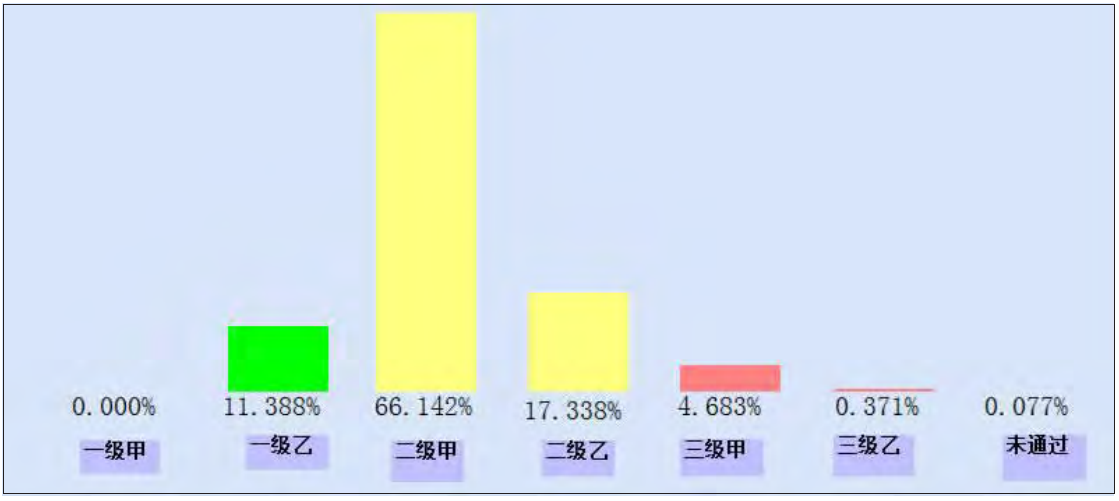


图 2 普通话测试等级统计图

图 2 是石家庄铁道大学 2011 年学生普通话测试等级统计图,各个等级的情况一目了然,对总体了解一个单位的普通话等级情况非常方便。

### (二)不同年份、不同单位测试成绩的比较

这个功能可以为我们提供横向和纵向的普通话测试成绩比较数据,对理性认识每一年的普通话成绩提供翔实数据。将石家庄铁道大学 2011 至 2013 年三年的"命题说话"项成绩做了比较,发现 2012 年的成绩明显高于 2013 年和 2011 年,而 2013 和 2011 年的成绩基本持平,我们觉得 2012 年学生的总体评测分数偏高了,而不是 2012 年那一批学生的普通话水平确实高过 2013 和 2011 年

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的学生。如果独立考察一年的测试成绩这样的问题是发现不了的。同理,不同学校的比较在这个程序中也可以比较,如果同层次学校同年级的学生普通话分数有较大差距,很可能是评分标准的把握上有问题。重要的是,有了这样的比较可以让我们发现更多问题,做更多的思考,使我们对普通话测试这个主观性较强的问题尽可能客观化。

### (三)得分异常情况处理

这项功能科大讯飞普通话智能测试系统也有,我们发现它还存在问题,一是统计有疏漏;二是有的异常检测不出。如前文提到的评分误差超过一定限度的有的就检测不出,可为疏漏问题。像有的扣分标准只能扣 0、1、2 分,结果扣了 3 分或 1.5 分这类情况科大讯飞系统是检测不出的。在我们这个系统里这类问题都得到了较好的解决。

## 八、结语

计算机辅助普通话水平测试“命题说话”项人工评分误差控制系统总体而言是比较科学的,它能及时监控到评测教师的评测误差,将问题控制在开始阶段;还有就是其客观性,发现问题依据的是具体的数据而不是凭感觉,更具有说服力;蔡伟在谈到高考作文评分误差控制时说:“参照量表最大的也是公认的优点是直观、具体、形象,更易使评卷员产生一致的理解,便于操作。评分标准和参照量表都是为了克服评分的随意性而编制的。”[6]这套系统的设计初衷也是用量化的标准来约束评测员打分的随意性。另外,在实时的检测过程中还能发现哪些测试员更负责,水平更高,相反水平低、不负责任的测试员也难逃这些数据的检测,为选拔一批优秀的测试员提供了较准确的数据。针对“命题说话”项主观性很强、容易出错的特点,尽快开发比较科学实用、具有更高水平的计算机辅助人工测试评分误差控制系统,将问题解决在过程或是萌芽之中,应是十分必要的。

## 参考文献:

[1]雷竣.计算机辅助普通话测试的问题思考及技术对策[J].武汉理工大学学报,2010(13):161-163.
[2]杨志明,张雷.改进普通话测试的概化理论分析[J].湖南师范大学教育科学学报,2003(1):76-82.
[3]陈佳民.用电脑建立常模监控高考作文评分[J].华南师范大学学报,1996(2):87-91.
[4]张雷,侯杰泰.普通话测试的录音评分可行性、信度及经济效率[J].心理学报,2001(2):97-103.
[5]丁琳.基于互联网网上阅卷系统的设计和实现[D].上海:华东师范大学,2003.
[6]蔡伟,娄庆华.高考作文评分误差控制构想[J].教育理论与实践,2008(7):24-25.

# The Preliminary Ideas on Proposition Speaking Error Control in Computer-aided Putonghua Proficiency Test

Yang Huiyong, Zhang Ning
(College of Humanities, Shijiazhuang Tiedao University, Shijiazhuang, 050043, China)

**Abstract**: The Proposition Speaking in Putonghua Proficiency Test has strong subjectivity. We design and develop a set of computer software which can be used to monitor the testers and control their errors to the minimum. The monitor system includes six aspects, i.e. error control, scoring consistency among testers, scoring consistency of individual tester, error control different test site, and the errors found by administrators.

**Key words**: Putonghua; Proposition Speaking; proficiency test; error control

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net