口语测试模糊评分方法设计及实验研究*

上海交通大学 金 檀 王 琰 宋春阳 郭曙纶

提要:本文对口语测试模糊评分方法进行设计及实验研究。在分析不同口语测试评分方法的基础上,本研究将语言测试理论与模糊控制理论相结合,提出了口语测试的三种模糊评分方法:(1)整体主观模糊评分法;(2)分项主观模糊评分推理法;(3)分项主观模糊评分加权法。采用这三种模糊评分方法对34名汉语言专业留学生进行实验,结果表明,三种方法所得的分数均服从正态分布,之间没有显著差异,并且与现行评分方法所得的口试分数、笔试分数之间分别呈显著相关及切实相关关系。

关键词:口语测试、模糊评分、评分方法

[中图分类号] H319 [文献标识码] A [文章编号] 1003-6105(2008)02-0157-08

1. 引言

1.1 选题

评分方法是测试过程中至关重要的环节(Bachman & Palmer 1996),在现行的口语测试评分方法中,评分员往往要根据整体或分项评分量表通过某一特定的数值对考生的表现进行评价(Underhill 1987; Fulcher 2003; Luoma 2004)。在实际的口语测试评分操作中,我们发现仅使用某一特定数值进行评价较为困难,因而本研究尝试设计口语测试模糊评分方法,先对考生的表现进行"范围"上的评价,而后推算出明确的"数值"。

1.2 解题

口语测试评分方法是评分员根据评分规则对考生表现进行评价的方法。根据评价的"前""中""后"三个阶段,评分方法可有如下分类:

(1)评价"前":根据评分规则(或评分员)对口语能力是否可分所持的观点(整体或分项),评分方法可分为整体评分法和分项评分法。

(2)评价"中":根据评分员评价的主观或客观性质,评分方法可分为主观评分法和客观评分法。

(3)评价"后":根据评分员给出的分数是"数值"还是"范围",评分方法可分为精确评分法和模糊评分法。关于模糊评分法,张文忠、郭晶晶(2002)提出了模糊评分这一思路,具有创新意义。

图1是口语测试评分方法分类图,由于现阶段采用真正意义上的客观评分并不现实并且本文主要研究模糊评分,因而本研究不考虑"客观评分"和"精确评分"。经过评价的"前""中""后"三个阶段(虚线框表示),模糊评分方法具体实现为整体主观模糊评分法和分项主观模糊评分法。本研究还对分项模糊评分采用推理和加权两种方法进行处理,因此,本研究中的口语测试模糊评分方法主要指以下三种:

(1) 整体主观模糊评分法;

(2) 分项主观模糊评分推理法;

(3) 分项主观模糊评分加权法。

1.3 研究问题

本研究中的模糊评分方法主要根据智能

* 感谢《现代外语》编辑部和匿名审稿专家提出的宝贵意见。感谢上海交通大学国际教育学院提供的帮助。本文受到上海交通大学人文社会科学基金项目"汉语水平机助自适应测试系统题库建设理论研究"(编号04-31)和上海交通大学文科科研创新基金项目"留学生汉语口语语料库建设及研究"(编号07QN008)的资助。

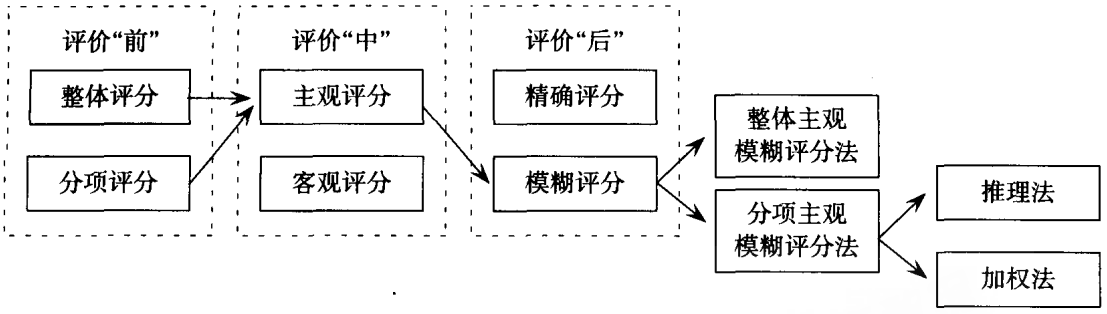

图 1 口语测试评分方法分类图

控制理论中模糊控制的原理设计而成。模糊控制是扎德在 1965 年首次提出的,它是一种应用模糊集合理论的控制方法,它不仅提出了一种用于实现基于知识(规则)甚至语义描述的控制规律,而且为非线性控制器提供了一个比较容易设计的方法(蔡自兴 1998)。本文尝试将模糊控制原理运用到口语测试评分中,旨在解决以下两个问题:

(1) 评分方法设计:三种模糊评分法具体如何设计?

(2) 评分方法效果:三种模糊评分法的实验结果如何?

2. 口语测试模糊评分方法设计

2.1 整体主观模糊评分法

整体主观模糊评分法是对考生表现的整体评价,由“模糊化”和“清晰化”两个步骤组成。“模糊化”过程是指评分员通过模糊集合对考生的表现打分。本研究将每个模糊集合用 5 个等级表示:不及格、及格、一般、良好、优秀。所有的模糊集合都用这 5 个元素及其隶属度来表示 (图 2 是隶属度函数示意图)。隶属度表示对应分数属于某个概念的程度,例如,某评分员对考生口语整体水平的 5 个模糊概念的隶属度进行打分:考生在 10%的程度上属于“不及格”概念,在 65%的程度上属于“及格”概念,在 20%的程度上属于“一般”概念,在 5%的程度上属于“良好”概念,在 0%的程度上属于“优秀”概念。其实,模糊概念的评分就是语言变量 T(整体评价)={不及格,及格,一般,良好,优秀}中每个元素的隶属度。应当指出,所有元素的隶属度之和须为定值,本研究中隶属度之和取 100,便于评分员理解和操作。同理,分项主观模糊评分即为 4 个考察项目(语音、内容、准确度和流利度)分别进行类似的操作。由于整体主观模糊评分和分项主观模糊评分的等级标准是一致的,因此两者的模糊集合隶属度函数图也相同。

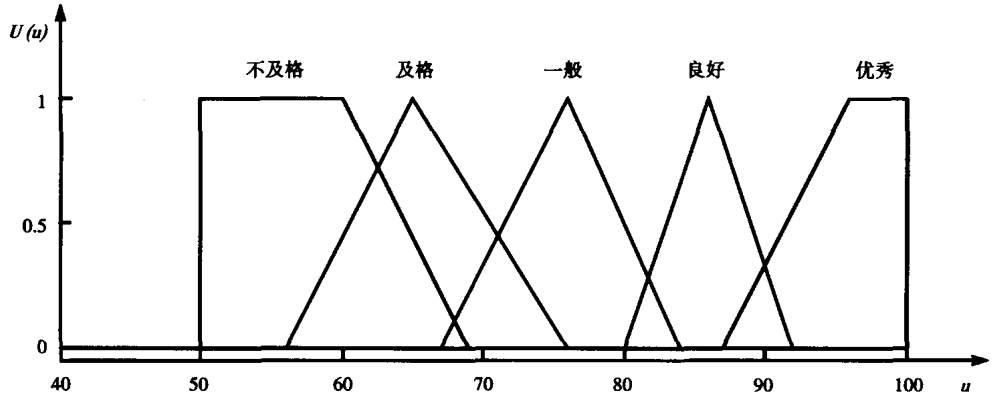

图 2 隶属度函数示意图

本研究的“清晰化”主要通过重心法进行，通过取模糊隶属度函数曲线同基础变量轴所围面积的重心的横坐标作为清晰值。例如，某评分员采用整体主观模糊评分法对某考生的表现进行整体评价，结果为{10，65，20，5，0}，化为标准隶属度表示为{0.10，0.65，0.20，0.05，0.00}，图 3 中的阴影部分就是这个隶属度函数曲线与基础变量轴所围成的图形，这个图形重心的横坐标就是清晰值，也就是该考生的精确分数。

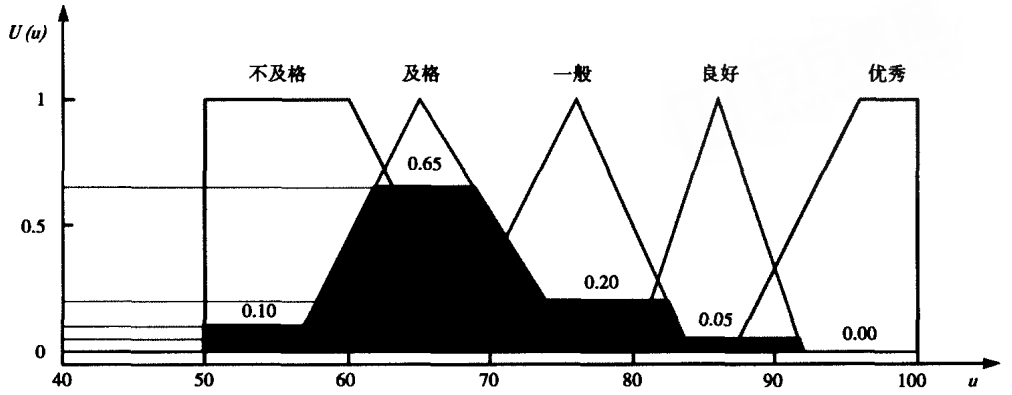

图 3 清晰化过程

2.2 分项主观模糊评分推理法

分项主观模糊评分推理法由“模糊化”、“模糊推理”和“清晰化”三个步骤组成。首先获取考生“语音”、“内容”、“准确度”和“流利度”四个分项的模糊集合分数，即“模糊化”过程(同 2.1 模糊化方法)。然后进行模糊推理，模糊推理是通过模糊逻辑理论和推理规则将多个模糊集合推理成一个模糊集合的过程，“它是在二值逻辑三段论的基础上发展起来的”(蔡自兴 1998：119)。本研究在推理过程中采用两级推理的方法(如图 4)：4 个输入先两两分别进行一级推理，其结果进行二级推理，然后再进行“清晰化”。

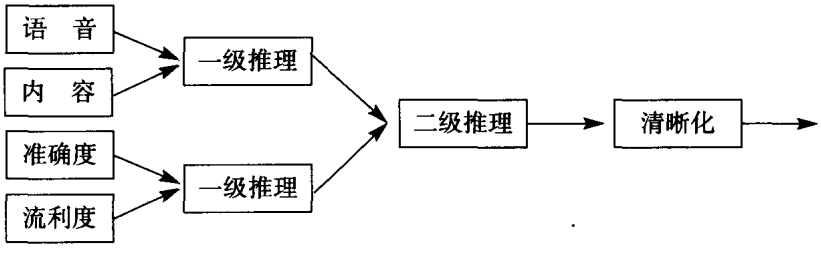

图 4 分项评分两级模糊推理

本文使用 Mamdani 直接推理算法进行推理(易继锴、侯媛彬 1999)。以下以某评分员对某考生“语音”和“内容”的评分为例进行说明。首先对表 1 中对应的两个模糊集合的隶属度两两进行取小操作，然后根据研究需要制定模糊推理规则(见表 2)，将表 1 中的结果对应到表 2 中可得到表 3。对表 3 中的 25 个结果按照所属的“不及格”、“及格”、“一般”、“良好”、“优秀”分别进行归类，可得：不及格={0}，及格={0,0,5,5,15,15,60}，一

般={0, 0, 5, 5, 5, 5, 20, 20, 30},良好={0, 0, 5, 5, 5},优秀 = {0, 0, 5}。然后对每个类别进行取大操作:max 不及格 = max {0} = 0,max 及格= max {0, 0, 5, 5, 15, 15, 60} = 60,max 一般 = max{0, 0, 5, 5, 5, 5, 20, 20, 30} = 30,max 良好 = max {0, 0, 5, 5, 5}=5,max 优秀 = max {0, 0, 5} = 5。因此,该考生语音和内容推理结果为{0, 60, 30, 5, 5}。"准确度"和"流利度"的推理同上述方法,再将两个推理的结果进行二级推理(同上述方法),得到的结果再进行"清晰化"(同 2.1 清晰化方法),可得到该考生的精确分数。

表 1 某考生语音和内容分数的推理(1)

		语音				
		不及格 0	及格 15	一般 60	良好 20	优秀 5
内容	不及格 5	0	5	5	5	5
	及格 60	0	15	60	20	5
	一般 30	0	15	30	20	5
	良好 5	0	5	5	5	5
	优秀 0	0	0	0	0	0

表 2 模糊推理规则表

	推理结果	输入变量 u_1				
		不及格	及格	一般	良好	优秀
输入变量 u_2	不及格	不及格	及格	及格	一般	一般
	及格	及格	及格	及格	一般	良好
	一般	及格	及格	一般	一般	良好
	良好	一般	一般	一般	良好	优秀
	优秀	一般	良好	良好	优秀	优秀

表 3 某考生语音和内容分数的推理(2)

	推理结果	输入变量 u_1									
		不及格		及格		一般		良好		优秀	
输入变量 u_2	不及格	不及格	0	及格	5	及格	5	一般	5	一般	5
	及格	及格	0	及格	15	及格	60	一般	20	良好	5
	一般	及格	0	及格	15	一般	30	一般	20	良好	5
	良好	一般	0	一般	5	一般	5	良好	5	优秀	5
	优秀	一般	0	良好	0	良好	0	优秀	0	优秀	0

2.3 分项主观模糊评分加权法

分项主观模糊评分加权法由"模糊化"、"加权"和"清晰化"三个步骤组成。首先获取考生"语音"、"内容"、"准确度"和"流利度"四个分项的模糊集合分数,即"模糊化"过程(同 2.1 模糊化方法)。然后将四项模糊分数按照设定的权系数(本研究中均取 1)进行加权,得到一个加权后的模糊集合后再进行"清晰化"(同 2.1 清晰化方法)。

以下用某评分员对某考生的评分为例进行说明(见表 4),加权后的模糊集合为{5, 38.75, 46.25, 8.75, 1.25},然后将该模糊集合进行清晰化,可得到该考生的精确分数。

表 4 某考生分项分数的加权

项 目	不及格	及格	一般	良好	优秀
语 音	0	15	60	20	5
内 容	5	60	30	5	0
准确度	10	40	40	10	0
流利度	5	40	55	0	0
加 权	(0+5+10+5)/4	(15+60+40+40)/4	(60+30+40+55)/4	(20+5+10+0)/4	(5+0+0+0)/4
(权系数均取 1)	= 5	= 38.75	= 46.25	= 8.75	= 1.25

3. 口语测试模糊评分方法实验

3.1 实验对象

本研究以上海某高校汉语言专业二年级留学生为实验对象(被试),共 34 人参加了本次口语测试实验,表 5 是被试的总体情况。

3.2 实验工具

(1) 测试工具:试题、评分标准及评分表[1],口语测试说明、准考证及考场情况记录表,录音设备。

(2) 数据处理工具:MATLAB 7.0 和 SPSS 13.0。

表 5 被试总体情况一览表

性别		国籍						
男	女	俄 国	韩 国	几内亚比绍	美 国	挪 威	日 本	印度尼西亚
17	17	1	17	1	2	1	11	1

3.3 实验过程

2007 年 7 月 9 日和 7 月 10 日实施口语测试,考生分别通过候考教室、准备教室及测试教室完成测试任务,两名评分员独立评分,测试后将相关材料存档,并完成评分数据的输入和校对工作。使用 MATLAB 7.0 为评分方法编写程序并获取实验结果,图 5 是程序流程图。

3.4 实验结果

实验得到以下结果(见表 6):(1)整体主观模糊评分法结果(记为“整体”);(2)分项主观模糊评分推理法结果 (记为 “推理”);(3)分项主观模糊评分加权法结果(记为“加权”)。

3.5 结果检验

3.5.1 正态性检验

对三种方法所得的实验结果通过 SPSS 13.0 进行正态性检验。表 7 正态性检验表中包括 Kolmogorov-Smirnov 检验和 Shapiro-Wilk 检验的结果。从表 7 中可知,这两项检验的结果都没有拒绝零假设,显著值均远远大于 0.05,因此,这三种方法所得分数的分布均服从正态分布。

3.5.2 单因素方差检验

这三种方法所得的分数是否存在显著差异?我们通过 SPSS 13.0 进行单因素方差检验来回答这个问题。表 8 是方差齐性检验表。表中显著值为 0.563,大于 0.05,表明各组方差是同质的,可以进行单因素方差检验。表 9

[1] 试题题型参照《高等学校外国留学生汉语言专业教学大纲》对二年级学生言语能力中“说”的要求(国家对外汉语教学领导小组办公室 2002: 13)设计;评分标准参照“口语考试五级标准”(北京语言大学汉语水平考试中心 2003: 14)及英语专业四级口试“评分标准”(文秋芳 1999: 106-107)制定了整体和分项评分标准。

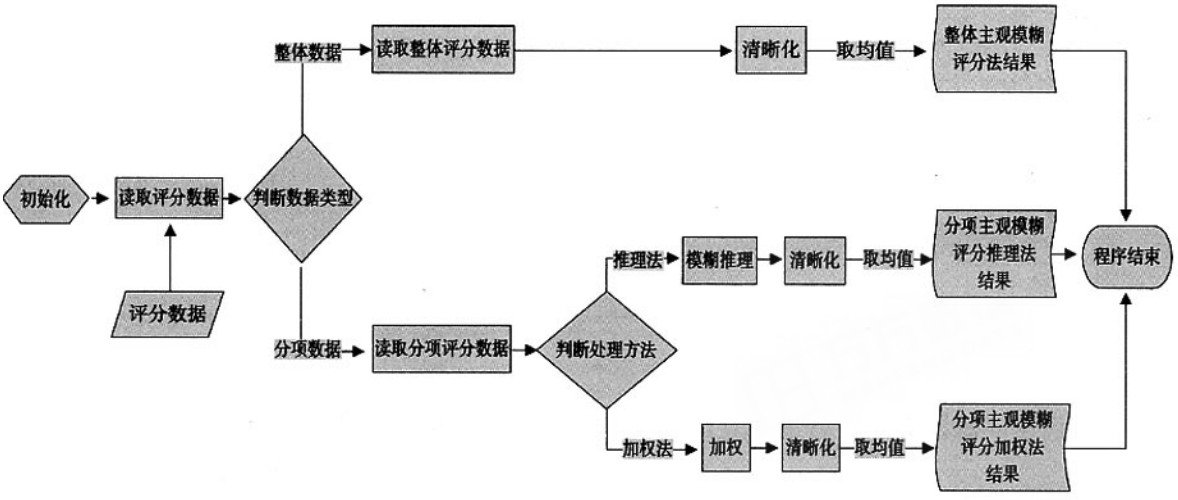

图 5 程序流程图

表 6 实验结果数据表

编号	整体	推理	加权	编号	整体	推理	加权
1	72.32	77.16	74.89	18	79.68	82.31	81.17
2	69.48	71.87	71.27	19	80.85	81.66	79.38
3	75.56	82.23	80.03	20	67.15	64.17	67.76
4	85.02	89.36	85.01	21	73.58	72.19	74.03
5	73.52	78.87	75.62	22	68.98	65.96	70.00
6	84.31	89.11	83.85	23	70.84	71.39	72.02
7	73.51	73.90	74.10	24	71.88	74.72	72.61
8	77.50	78.34	75.98	25	69.99	69.35	69.77
9	77.44	82.47	79.52	26	79.00	81.52	78.19
10	81.01	83.57	81.96	27	80.18	82.25	79.83
11	82.47	83.25	82.00	28	93.07	94.99	93.00
12	83.97	83.99	83.12	29	74.26	72.61	73.47
13	73.17	72.95	73.41	30	77.13	79.01	75.18
14	74.05	80.98	77.56	31	86.23	87.88	85.08
15	83.03	83.51	81.57	32	70.99	70.46	71.12
16	93.07	94.98	92.74	33	89.42	90.42	89.80
17	77.89	81.31	79.83	34	82.52	83.20	80.99

表 7 正态性检验表

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	统计量值	自由度	显著值	统计量值	自由度	显著值
整体主观模糊评分法	.124	34	.200*	.959	34	.233
分项主观模糊评分推理法	.122	34	.200*	.972	34	.504
分项主观模糊评分加权法	.091	34	.200*	.960	34	.246

*真显著性的下限。 a Lilliefors 显著性校正。

是单因素方差检验表。表中显著值为0.555,大于0.05。因此我们可以得出结论,三种模糊评分方法所得的分数之间不存在显著差异,即三种不同的方法对评分结果没有显著影响。

表8 方差齐性检验表

Levene 统计量	自由度1	自由度2	显著值
.578	2	99	.563

表9 单因素方差检验表

	平方和	自由度	均方	*F* 值	显著值
组间方差	56.398	2	28.199	.592	.555
组内方差	4717.975	99	47.656		
总和	4774.374	101			

3.5.3 相关分析

在语言测试中，四项技能存在着实质上的相关联系(Wood 1993)[2]。本研究选取“汉语视听”、“汉语口语”、“汉语阅读”和“汉语写作”分数(采用现行评分方法)与三种模糊评分法所得的分数进行相关分析。应当指出,本实验的口语测试于2007年7月9日和7月10日进行,该批考生在2007年7月9日和7月11日还参加了“汉语视听”、“汉语写作”和“汉语阅读”的期末笔试,具有较好的共时效度;同时考虑到本次实验是“汉语口语”的期末考试,因而还选取了该批学生“汉语口语”的期中考试分数(采用现行评分方法,以下记为“汉语口语”)进行相关分析。

表10是相关分析数据，从表中可知,三种模糊评分方法所得的分数与“汉语口语”分数的相关系数均在0.01水平上显著,分别为0.806、0.761、0.799,平均值为0.789(保留小数点后3位)，与笔试分数的相关系数均在0.01或0.05水平上显著,最低为0.398,最高为0.505，平均值为0.459（保留小数点后3位)。

表10 相关分析数据

		口试	笔试		
		汉语口语	汉语视听	汉语阅读	汉语写作
整体主观	皮尔逊积矩相关	.806**	.502**	.505**	.409*
模糊评分法	显著值(双尾检验)	.000	.003	.002	.016
	样本容量	34	32	34	34
分项主观	皮尔逊积矩相关	.761**	.398*	.458**	.424*
模糊评分推理法	显著值(双尾检验)	.000	.024	.006	.012
	样本容量	34	32	34	34
分项主观	皮尔逊积矩相关	.799**	.490**	.499**	.447**
模糊评分加权法	显著值(双尾检验)	.000	.004	.003	.008
	样本容量	34	32	34	34

注:**表示相关在0.01水平上显著(双尾检验),*表示相关在0.05水平上显著(双尾检验)。

[2] 笔者认为,在汉语作为第二语言的测试研究中,由于汉字的因素以及学生第一语言背景的不同(例如日韩学生的汉字认读能力较强、欧美学生口语表达能力较强等),相关联系程度的大小也不一样。

4. 结论

在第二语言口语测试实际的评分过程中，仅用某一特定的数值来评价考生的表现较为困难，也存在着一定的问题。本文尝试先对考生的口语表现进行“范围”的评价而后进行特定“数值”的推算，以期提高口语测试评分质量。口语测试模糊评分法是将模糊控制理论应用于分数评定的一种方法，本文对三种模糊评分方法进行了初步设计和实验研究。通过对实验结果的检验，得到以下结论：

(1) 正态性检验的显著值均远远大于0.05，三种模糊评分方法所得的分数均服从正态分布。

(2) 方差检验的显著值为0.555，三种模糊评分方法所得的分数之间没有显著差异。

(3) 三种模糊评分方法所得的分数与现行评分方法所得的口试分数、笔试分数之间分别呈显著相关(平均值0.789)及切实相关(平均值0.459)关系。

应当指出，本研究中的隶属度函数、推理规则和权系数均为实验设定，如果能对大规模的口语测试结果进行统计分析并以某项指标进行最优化研究，则有望得出更为合理、科学的隶属度函数、推理规则以及权系数，从而进一步发展和完善模糊评分法在口语测试中的应用。

参考文献

Bachman, L. F. & A. S. Palmer. 1996. *Language Testing in Practice* [M]. Oxford: Oxford University Press.

Fulcher, G. 2003. *Testing Second Language Speaking* [M]. London: Pearson Education Limited.

Luoma, S. 2004. *Assessing Speaking* [M]. Cambridge: Cambridge University Press.

Underhill, N. 1987. *Testing Spoken Language: A Handbook of Oral Testing Techniques* [M]. Cambridge: Cambridge University Press.

Wood, R. 1993. *Assessment and Testing: A Survey of Research* [M]. Cambridge: Cambridge University Press.

北京语言大学汉语水平考试中心，2003，中国汉语水平考试大纲·高等 [M]。北京：北京语言大学出版社。

蔡自兴，1998，智能控制——基础与应用 [M]。北京：国防工业出版。

国家对外汉语教学领导小组办公室，2002，高等学校外国留学生汉语言专业教学大纲 [M]。北京：北京语言大学出版社。

文秋芳，1999，英语口语测试与教学 [M]。上海：上海外语教育出版社。

易继锴、侯媛彬，1999，智能控制技术 [M]。北京：北京工业大学出版社。

张文忠、郭晶晶，2002，模糊评分：外语口语测试评分新思路 [J]。现代外语(1)：98-102。

收稿日期：2007-09-02；
作者修改稿，2007-11-19；
本刊修订，2008-03-25

通讯地址：200030 上海交通大学国际教育学院(金)
<sandjintan@yahoo.com.cn>
200240 上海交通大学自动化系(王)
<yanw@sjtu.edu.cn>
200030 上海交通大学国际教育学院(宋)
<songchy@sjtu.edu.cn>
200030 上海交通大学国际教育学院(郭)
<gshulun@163.com>