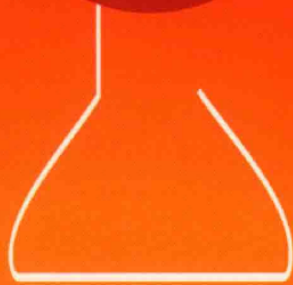


YUYAN  
CESHI QIUSUO

语言  
测试  
求索



邹

申

著

湖南师范大学出版社

YUYAN  
CESHI QIUSUO

# 语言 测试 求索

邹  
申  
著

湖南师范大学出版社

## 图书在版编目(CIP)数据

语言测试求索 / 邹申著. —长沙:湖南师范大学出版社, 2018. 6  
ISBN 978-7-5648-3238-4

I. ①语… II. ①邹… III. ①英语—测试—文集 IV. ①H319-53

中国版本图书馆 CIP 数据核字(2018)第 096452 号

## 语言测试求索

邹 申 著

◇组稿编辑:李 阳

◇责任编辑:李健宁 江洪波

◇责任校对:李 航

◇出版发行:湖南师范大学出版社

地址/长沙市岳麓山 邮编/410081

电话/0731-88873071 88873070 传真/0731-88872636

网址/<http://press.hunnu.edu.cn>

◇经销:新华书店

◇印刷:长沙印通印刷有限公司

◇开本:710mm×1000mm 1/16

◇印张:18

◇字数:380千字

◇版次:2018年6月第1版 2018年6月第1次印刷

◇书号:ISBN 978-7-5648-3238-4

◇定价:68.00元

凡购本书,如有缺页、倒页、脱页,由本社发行部调换。

本社购书热线:0731-88872256 88873071

投稿热线:0731-88872256 13975805626 QQ:1349748847

## 序

外语教学与外语测评密不可分。教学的结果应通过测评得以检验，而测评的结果则应进一步推动教学的进步与发展。因此，外语测评是外语教学的落脚点，亦是新的出发点。

由邹申教授编撰的《语言测试求索》一书共分为四个部分。全书从理论和实践的多个角度展开，围绕考试大纲的研究，英语专业四、八级考试的研究，语言测试的研究方法以及英语能力等级量表的研究等诸多主题进行探讨。全书以英语专业四、八级考试为主线，贯穿了我国语言测试研究领域在过去二十余年间的发展历程。虽然本书的研究成果在整个语言测试学界只是一个缩影，但却实实在在凸显了作者在语言测试研究中不断积累、不断思考、不断求索、不断创新的精神。

邹申教授是长期活跃在我国语言测试领域的著名专家，特别是对高等学校英语专业四、八级考试倾注了毕生的心血与精力。因此，本书的研究成果一定会对英语专业四、八级考试的健康可持续发展发挥积极作用。此外，我也欣喜地看到，邹申教授先后培养了近六十位博士（后），其中许多人活跃在语言测试研究领域并已在学术上崭露头角。因此，我也相信，本书的出版将对我国语言测试研究产生一定影响。

是为序。

戴炜栋

教育部社科委员会语言文学学部召集人

上海外国语大学原党委书记、校长

教育部高等学校外语专业教学指导委员会原主任委员

国务院学位委员会外语学科评议组原组长

# 目 录

## 第一部分 理论求索——考试大纲研究

考试、教学及其他

——写在《英语专业四、八级考试大纲》出版之际 ..... (003)

对标准参照考试的再认识

——兼谈其在考试反拨效应中的积极作用 ..... (009)

语言教学大纲与语言测试的衔接

——TEM8 的设计与实施 ..... (020)

关于考试科学属性的思考

——兼谈高校英语专业四、八级考试大纲(2004 年新版)的修订 ..... (029)

对考试效应的认识与对策

——兼谈高校英语专业四、八级考试大纲的修订原则与方案 ..... (038)

试论口语测试的真实性 ..... (047)

回顾与展望

——写在英语专业四、八级考试开考 20 周年之际 ..... (054)

## 第二部分 实践求索(上篇)——英语专业四、八级考试研究

听力测试中交互性的探究

——兼谈 TEM8 考试听力项目的修订 ..... (069)

阅读测试中题目类型、策略与分数的关系

——TEM4 考试阅读项目的答题效度研究 ..... (079)

TEM8 写作能力评估

——要求、问题及对策 ..... (088)

他们如何使用写作评分标准?

——TEM4 新老评分员调查 ..... (095)

写作测试中写作环境信息对构思阶段的影响 ——一项基于大三学生的调查 .....	(104)
A Washback Study of the Test for English Majors for Grade Eight (TEM8) in China—From the Perspective of University Program Administrators .....	(113)
提升考试公平性 ——以英语专业四、六级考试为例 .....	(141)

### 第三部分 实践求索(下篇)——语言测试的研究方法

应用 EQS 和 BILOG 分析 TEM8 人文知识项目的构念效度 .....	(153)
基于多层面 Rasch 模型的 TEM4 作文评分员效应研究 .....	(165)
TEM4 评分效度与计算机辅助评卷 .....	(179)
语料库在试题设计和验证中的应用研究 .....	(188)
基于语料库的写作测试效度研究 ——以《英语专业写作教学语料库》为例 .....	(198)
民族志研究方法在语言测试研究中的应用 .....	(209)
考试评估中的信息反馈 ——谈间接数据的应用 .....	(219)

### 第四部分 求索进行时——英语能力等级量表研究

《欧洲语言共同参考框架》在中国:研究现状与应用展望 .....	(229)
Exploring the Adaptability of the CEFR in the Construction of a Writing Ability Scale for Test for English Majors .....	(240)
参考文献 .....	(263)
后记 .....	(282)

## 考试、教学及其他

# 理论求索——考试大纲研究

### 第一部分



考试是一种测量工具,而考试大纲则是如何设计测量工具的重要图纸。我国英语专业四、八级考试已经走过了二十余年,在其开发、效度验证、结构调整和维护过程中,不同版本的考试大纲所起到的作用可见一斑。本书的第一部分以理论的探索为主线,从考试大纲的研究出发,说明如何在考试大纲的设计中体现语言测试理论前沿的最新研究成果,并渗透到施考的各个环节之中。

这一部分共收录了7篇代表性论文。第1篇论文主要论述了《英语专业四、八级考试大纲》(1994年版)的起草、论证、出台以及实施的各个细节。第2篇论文通过剖析标准参照考试的实质与属性,对英语专业四、八级考试的性质、定位和反拨效应等提出了深刻的见解。第3篇论文梳理了语言教学大纲与语言测试的关系,提出英语专业八级考试的特征及其对英语专业教学的意义。第4篇论文结合《英语专业四、八级考试大纲》(2004年版),提出了如何有效提高考试科学性的途径,以此丰富考试科学性的内涵。第5篇论文从考试效应出发,具体阐述了《英语专业四、八级考试大纲》(2004年版)的理论依据和实践操作,并说明了英语专业四、八级考试在结构上的调整。第6篇论文从口语测试真实性的角度出发,提出了口语测试任务与考生之间的交互关系。第7篇论文回顾并总结了英语专业四、八级考试所经历的二十年,并对该考试的将来进行了展望和探讨。



## 考试、教学及其他

——写在《英语专业四、八级考试大纲》出版之际

《英语专业四八级考试大纲》(以下简称《考纲》)的起草工作于1992年底开始。在《考纲》的设计过程中制订小组成员通过座谈及问卷调查的形式广泛征求了有关人士对四、八级考试的意见及建议。小组成员也曾深入讨论了《英语专业基础阶段教学大纲》和《英语专业高年级教学大纲》,力求领会两个教学大纲的基本原则与精神。与此同时《考纲》制订小组也参阅了国内外英语测试方面的有关论著使《考纲》的基本思路与框架更趋于科学化、合理化。在国家教委和外语专业教学指导委员会英语组有关领导、专家的直接关心下,经过全体有关人员的努力,《考纲》已于1994年初由上海外语教育出版社正式出版发行。今年英语专业四、八级考试已分别按照《考纲》的要求进行命题、编排试卷和组织实施。预计在今后的几年内,英语四、八级考试均将根据《考纲》所订的标准举行。可以这样说,《考纲》问世的意义不在其本身,而在它对今后英语专业学生考试乃至全国英语专业教学将要产生的影响。本文将对此谈一些笔者的看法。

### 1. 为何要制订《考纲》

考试历来是各级各类教学中一个必不可少的环节,英语专业的教学亦无例外。目前,各校英语专业都有自己的课程设置并在此基础上规定不同的教学评估形式。一般采用的有以下几种:新生入学的分班考试、期中/末的课程考试、决定升留级的水平考试。可能有人会问:既然各校英语专业都有自己的一系列教学检查措施,为何还要制订《考纲》呢?我们的认识如下。

凝聚着众多专家学者心血及智慧的《高等学校英语专业基础阶段教学大纲》和《高等学校英语专业高年级教学大纲》,在广泛调查、科学论证的基础上

完稿,并先后于1989年与1990年出版。至今这两份教学大纲已实施多年。教学大纲的付诸实施标志着具有中国特色的英语专业教学体系的日臻成熟与完善。教学大纲对教学原则、对象、任务与目的、内容、教学评估形式等都作了明确规定。另外,教学大纲亦对每一级学生在语言基本知识、语言单项与综合技能、交际能力、文化知识等诸方面应达到的水准提出了具体要求。教学大纲既考虑到我国英语专业教学和学生的实际情况,又吸取了国内外各种教学理论及教学法的精华。在这点上,教学大纲的出版使用对高等学校英语专业教学向更高层次发展有着深远的意义。

由于种种因素,各高等院校的英语专业在许多方面(比如课程设置、教材使用、培养目标、教学评估)存在着差异。在教学大纲未实施前人们很难用较统一的尺度衡量各校的英语专业教学。教学大纲的颁布弥补了这方面的缺陷,它给各校组织教学、开设课程、制订培养目标,提供了依据,可以使教学及课程设置正规化、合理化。此外,教学大纲对英语专业各类教材的编写具有较高的参考价值,可以使教材更具有针对性、可行性。同时,教学大纲给校内或校际的教学评估提供了同一套衡量标准,使教学评估具有统一性、有效性。这一切都有助于缩小校与校之间在上述几个方面存在的距离,推动英语专业教学。

两份教学大纲经过一段时期的贯彻使用后便需要检查各校的执行情况。在采用何种评估方式上,基础阶段教学大纲有明确的规定:即“通过统一测试的方法检查执行教学大纲的情况”(基础阶段教学大纲,1989年);统一测试的目的是“推动校际交流”,提供“改进教学的依据”(高年级教学大纲,1990年)。因此,高校英语专业有关专家、教授在几年前便开始了这方面的探索与实践。经过几年的努力,英语专业教学的评估形式——四、八级考试——已初具规模。

为了使考试更科学、更有效地检查基础阶段及高年级英语专业学生的英语水平,并使考试具有稳定性,其中一个重要环节就是考试大纲的制订:即规定考试目的、性质、范围、要求、题型、时间、分数计算等。《考纲》可以使两份教学大纲提出的教学评估具体化,这对有关人员及学校都是大有裨益的。从命题、试题审定人员的角度来看,《考纲》有助于他们把握题目的难度以及具体命题要求。从学校的角度来看,以《考纲》而不是某年的试卷作为教学参考,有利于从宏观上调整教学结构,以改进教学。总而言之,作为教学大纲的配套文件,《考纲》有助于客观地、公正地检查各校执行教学大纲的情况,以进一步推

动英语专业教学改革。

## 2. 教学与考试的关系

受英语组的委托,我们曾在部分高校英语专业二、四年级任课教师中进行了一次问卷调查(问卷调查结果将另行发表)。其中一项内容是征询四、八级考试是否有必要举行。在收回的问卷中 85%左右的教师认为“很有必要”、“较有必要”或“基本必要”。10%左右的教师认为“没有必要”,其理由主要有:(1)干扰正常教学秩序;(2)增加学生的额外负担;(3)考前搞突击强化训练;(4)各类学校的培养目标不同,对专业训练作用不大;等等。除第四条理由外(它牵涉到考试的效度问题),前三条理由比较集中地反映了教师中存在的一种忧虑。尽管这部分人不多,但他们的看法有一定的代表性,值得引起有关人士的重视。

在实行考试的过程中,我们应自始至终地摆正教学与考试的位置。我们必须清醒地认识到两者之间实际上存在着一种主次关系。教学是一个复杂的系统工程,考试只不过是这个过程中采用的一种辅助手段。其目的是为了检验教学质量。换句话说,测试应为教学服务,而不是教学围着测试转。所以,为了提高考试的平均分或及格率而打乱正常教学秩序,大搞“题海”战术的做法是不足取的。这种权宜之计或许会短期见效。但从长远来看,这对英语专业的发展及学生的培养都是弊多利少。我们当前英语专业的课程设置旨在全方位地培养学生,既开设语言实践课,又设有语言理论课;既有语言文学方面的课程,又有文化方面的课程。如果为了应付考试而将有关课程压缩甚至砍掉,那么我们高等院校全日制英语专业本科的课程与社会上五花八门的“应试班”就相差无几了。对学生本人来说,把许多精力放在应付考试上就意味着失去许多本应掌握和巩固语言运用以及社会文化基本知识 with 技能的机会。这样的学生是达不到教学大纲所规定的培养目标的。

在正确处理教学与考试关系的同时,我们不得不承认测试对教学影响的客观存在。在这方面,国外教育界及英语教学界已有许多专门的论著。但考试对教学的影响是积极的还是消极的,专家权威也是众说纷纭,争论不休。有的从正反两方面探讨了考试的效应问题,比如 Alderson 和 Wall(1993)提出了十五种考试效应假设。有的仅仅是承认这个现实,比如 Pearson(1988)认为“人们普遍接受这一点:即公开考试影响教师、学生以及家长的态度、表现以及动机”。有的持肯定的观点,比如 Morris(1972)认为“为保证课程设置的贯

彻执行,考试是有必要的”。Swain(1985)建议考试设计人员“努力使考试产生积极效应”。也有的抱否定的态度,比如,Vernon(1956)认为“考试歪曲了课程设置”。他的理由是教师往往忽略那些不直接影响考试成绩的课程及教学活动,并且致力于应试准备。总之,不管有关人士对考试的效应予以肯定还是否定,这种现象将随着考试的存在而继续下去。

当我们清楚地意识到考试具有导向性时,我们就应该采取一些方法使考试为教学提供正面的效应。

(1)测试内容要多层次化。由于英语专业四、八级考试的主要目的是检查二和四年级学生是否分别达到教学大纲所规定的教学标准,考试内容的确定主要取决于教学大纲所制订的教学要求。同时,考试内容的选择力求具有代表性。也就是说,测试内容既要包括单项语言技能的考核,也要兼顾语言综合运用能力的评估。为了使考试真正达到教学评估的目的,在上述两者间的比例编排上,考试也应体现出教学大纲对不同阶段学生的英语水平的不同要求。

在测试某一单项语言技能时,考试应区分其高层次与低层次的次技能并根据教学大纲的要求,确定这些次技能在某一测试项目中的比例。例如,四级考试中阅读理解 Section A 主要由两个部分构成:①了解主旨大意和重要细节;②语言知识的使用(Weir, 1992)。相比之下,其中①属于较高层次的阅读技巧。同时这两个部分又可再划分成以下几个小部分(Weir, 1992):

#### ①了解主旨大意和重要细节

##### a) 获取主旨大意及重要细节

——归纳主旨大意或重要细节

——区分篇章中不同的观点(例如:有关与无关节节)

##### b) 区别事实与观点、了解作者的态度等

##### c) 理解隐含意义

——通过背景知识理解篇章

——通过上下文或字里行间的关系理解篇章

#### ②语言知识的使用

##### a) 理解语法概念(比如:因果关系、目的、对比等)

##### b) 理解句子与从句的语法结构

##### c) 猜测生字及陌生词组的含义

##### d) 识别篇章内的关联词以及前后呼应的手段

根据四级教学大纲的要求,阅读能力的测试重点应在①上。而在①中间,a)、b)和c)三个构成因素都应在具体题目中有所体现。但比例的掌握要适当,以避免题目过难或过易。总而言之,对测试内容进行层次化处理,我们既可以减少测试的盲目性,又可以避免出现某一测试点的简单重复。同样重要的是阅读部分所采用的不同层次、不同种类的题目可以促使学校教师在日常教学过程中注意训练不同的阅读技能。

(2)测试形式要多样化。在选择测试形式的过程中,我们既要考虑到英语专业教学的特点、培养目标,又要考虑到考试的可行性及可操作性。当前,英语专业四、八级考试已初具规模,且考生人数在逐年增加。为了保证考试的分数具有一定的可靠性,有的测试项目采用了多项选择题的形式,便于考试后期的评分及分析工作的开展。然而,多项选择题也有其薄弱的一面。它属于分离式测试法的范畴。其特点是一道试题孤立地测试一个语言因素。因此,它无法对考生的语言综合运用能力进行较全面的评估。为弥补这点,其余测试项目采用主观题或综合测试法的形式,以提高试卷的效度。采取两者结合的方法的好处是可以使不同的测试形式取长补短,使考试更有效地检查教学质量。同时,我们也可避免由于偏重某一测试形式而对教学产生的不利因素。

(3)具体项目要有灵活性。为了能较全面地考核四、八级阶段学生的英语水平,我们在具体项目设计上要增加灵活性。根据教学大纲的要求,试题设计人员可就每一单项设计几种不同的试题形式,考纲的形式公布。以后命题时由命题人员从中选择。四、八级考试大纲已在这方面做了初步的尝试,目的是通过扩大命题的灵活性来减少应试技巧训练对考生成绩的作用。

### 3. 今后工作的几点设想

1994年英语专业四、八级考试已于上半年顺利举行。在回顾总结今年考试的基础上,我们看到了已取得的成绩,同时也找到存在的差距。今后测试工作需要从以下几个方面加强:

(1)及时从各种渠道(包括考生、教师、校方)收集对考试的反馈;加强校与校之间的沟通。

(2)进一步加强试测工作,以保证试卷的整体质量及试卷间的相对难度。

(3)逐步建立一支相对稳定的命题人员队伍,提高命题水平。

(4)加强考后统计分析工作;分析结果应尽快反馈到学校,供学校作教学

上的参考。

(5) 进一步改进主观题,尤其是写作与翻译的评分标准及评分工作,使之尽可能客观、公正地反映考生的真实水平。

测试是一个复杂的系统工程,需要有关人员的精诚合作与不懈努力。相信通过不断地探索,英语专业四、八级考试将更完善,更具有权威性。

(作者:邹申。原载于《外语界》1994年第3期)

## 对标准参照考试的再认识

### ——兼谈其在考试反拨效应中的积极作用

#### 1. 标准参照考试定义

教育测量界(包括语言测试界)对标准参照考试的关注和讨论由来已久。标准参照考试概念的提出最早可以追溯到 Glaser 和 Klaus(1962: 419-474)、Glaser(1963: 519-521)。Glaser 指出(1963: 519):

Criterion-referenced measures indicate the content of the behavioural repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a certain criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others.

Glaser 对标准参照考试的目的、性质和功能作了概括性的定义,对以后学界在这方面的进一步研究奠定了基础。

纵观语言测试领域内的一些较有影响的书籍,凡是介绍语言测试类别时,无不例外地会提及标准参照考试。比如, Henning(1987: 7)认为:

Characteristically criterion-referenced tests are devised before the instruction itself is designed. The test must match teaching objectives perfectly... Students are not evaluated by comparison with the

achievement of other students, but instead their achievement is measured with respect to the degree of their learning or mastery of the pre-specified content domain.

Hughes(1989: 18)指出:

The purpose of criterion-referenced tests is to classify people according to whether or not they are able to perform some task or set of tasks satisfactorily... This means that students are encouraged to measure their progress in relation to meaningful criteria, ...

Alderson 等人(1995: 287)提出:

If a test is criterion-referenced, students are compared not with one another, but with a level of achievement, or a set of criteria set out in marking descriptors.

Bachman(1990: 74)认为:

Criterion-referenced tests are designed to enable the test user to interpret a test score with reference to a criterion level of ability or domain of content.

Brown 和 Hudson(2002: 5)提出:

... a *criterion-referenced test* will be defined as any test that is primarily designed to describe the performances of examinees in terms of the amount that they know of a specific domain of knowledge or set of objectives.

综上所述,该类考试须建立在某一(系列)事先制订的标准(criterion)之上。该(系列)标准源自于课程教学目标(teaching objectives)、预期达到的学



习成就水平(levels of achievement)等。标准参照考试的另一重要特征是对成绩的解释方式。在标准参照考试中,我们对某一分数的解释不是基于该分数在同一考生群体中的相对位置之上,而是根据事先制订的目标等级标准,来解释分数的实际含义。换句话说,我们关注学生达到还是没有达到预先设定的水平,或达到程度如何。

## 2. 考试反拨效应及研究现状

本节主要探讨考试反拨效应以及相关研究。什么是考试反拨效应?如何认识其性质和作用? Hughes(1989: 1-2)认为,考试对教学和学习产生的作用可称为反拨效应。根据考试对教学和学习产生的积极或消极影响,考试反拨效应可分为正面效应或反面效应。正面效应对课程的完善、教材或教法的更新起到积极的推动作用;而负面效应不仅不对教与学起到推动作用,而且还可能会使教与学走入歧途。Hughes对考试反拨效应的讨论主要集中在对教学的影响程度上。

Bachman和Palmer(1996: 29-35)提出的测试影响度框架超越教学范围,属于宽泛型框架。根据他俩的观点,测试影响度指考试对社会、教育体系以及处于这些体系中的个体(包括考生、教师)所产生的影响。这种影响在两个层面上出现:微观及宏观层面。微观层面主要涉及受某一考试使用影响的个体。受宏观层面影响的主要是教育体系或社会。在Bachman和Palmer的框架中,测试对教学的影响——反拨效应——是作为测试影响度的一个方面。考试反拨效应不是一种孤立的现象,对教学的影响也会扩展到教育体系甚至于社会之中。

Brown和Hudson(2002: 47-48)则从课程开发与设置的角度,界定考试反拨效应的作用。根据他俩的观点,考试反拨效应通常指某一考试对某一阶段课程设置的影响;同时考试反拨效应也贯穿于整个课程设置过程,即会对整个课程体系产生影响。

考试反拨效应是否存在?对教学到底有没有影响?又在哪些方面体现出来?在语言测试领域最早从事反拨效应研究的是Alderson和Wall(1993: 115-129)。他俩在文中讨论了反拨效应的概念,提出了一系列反拨效应假设,回顾了这方面的研究;文章最后提出一系列研究设想,目的是进一步探索反拨效应的性质和表现。迄今为止,这方面的研究成果表明,考试反拨效应确实存在,而且其复杂性远远超出人们的想象(Brown, 2005: 243)。Alderson和