

中国汉语水平考试 (HSK) 研究书系

漢

HSK

语言测试专业
硕士论文精选

北京语言大学汉语水平考试中心 编

主编：孙德金

Selected Theses of Master Candidates
in Language Testing

北京语言大学出版社



中国汉语水平考试 (HSK) 研究书系

A Series of Studies on the Chinese Proficiency Test (HSK) of China

HSK

语言测试专业硕士论文精选

Selected Theses of Master Candidates in Language Testing

北京语言大学汉语水平考试中心 编

主编 孙德金

北京语言大学出版社

(京)新登字157号

图书在版编目(CIP)数据

语言测试专业硕士论文精选 / 北京语言大学汉语水平考试中心编, 孙德金主编.

—北京: 北京语言大学出版社, 2005

(中国汉语水平考试(HSK)研究书系)

ISBN 7-5619-1457-1

I. 中…

II. ① 北… ② 孙…

III. 汉语—对外汉语教学—水平考试—研究—文集

IV. H195-53

中国版本图书馆CIP数据核字(2005)第067059号

书 名: 语言测试专业硕士论文精选

责任印制: 汪学发

出版发行: **北京语言大学出版社**

社 址: 北京海淀区学院路15号 邮政编码: 100083

网 址: www.blcup.com

电 话: 发行部 82303648/3591/3651

编辑部 82303647

读者服务部 82303653/3908

印 刷: 北京北林印刷厂

经 销: 全国新华书店

版 次: 2005年7月第1版 2005年7月第1次印刷

开 本: 710毫米×1000毫米 1/16 印张: 17.125

字 数: 237千字 印数: 1-2000

书 号: ISBN 7-5619-1457-1/H.05069

定 价: 39.00 元

凡有印装质量问题, 本社负责调换, 电话: 82303590

中国汉语水平考试 (HSK) 研究书系

第一辑

《语言测试及测量理论研究》

主编：张 凯

《中国汉语水平考试 (HSK) 研究报告精选》

主编：谢小庆

《语言测试专业硕士论文精选》

主编：孙建奎

《中国汉语水平考试 (HSK) 技术报告》

主编：郭树军

《中国汉语水平考试 (HSK) 工作规范》

主编：陈 宏

A Series of Studies on the Chinese
Proficiency Test (HSK) of China

Collection One

*Studies on Language Testing and Measurement
Theories*

Zhang Kai (ed.)

*Selected Research Reports on the Chinese
Proficiency Test (HSK) of China*

Xie Xiaoqing (ed.)

*Selected Theses of Master Candidates in Language
Testing*

Sun Dejin (ed.)

*Technical Reports on the Chinese Proficiency Test
(HSK) of China*

Guo Shujun (ed.)

*The Standards and Regulations of the Chinese
Proficiency Test (HSK) of China*

Chen Hong (ed.)

中国汉语水平考试（HSK）研究书系

编辑委员会

主 任

孙德金

委 员

陈 宏 郭树军 谢小庆 张 凯

中国汉语水平考试(HSK)自1984年在北京语言学院(今北京语言大学)开始研制,已经走过了二十多年的历程。她伴随着中国改革开放的脚步,伴随着对外汉语教学事业的发展,从无到有,从小到大,成为在国内外有广泛影响的汉语水平测量工具。今天我们编辑出版“研究书系”第一辑五本著作,既是对过去的梳理和总结,更是为了开拓更加广阔的未来,因为我们知道,汉语教学和汉语水平测试在中国国力不断增强的大背景下拥有可以预见的美好前景。

对外汉语教学,或称汉语作为第二语言教学,无论是作为一项事业还是作为一门学科,相对于英语来说,历史是短暂的。而作为对外汉语教学四大环节之一的汉语测试,其开发、研究的历史就更加短了。上世纪80年代初,国内以英语为代表的第二语言教学开始由重知识转向强调语言运用能力,英语的标准化考试 TOEFL 对听力和阅读能力的考查对英语教学思想的改变产生了重要作用。而在对外汉语教学界,人们当时关注的只是教学中的课程测试或称成绩测试,还没有一种类似 TOEFL 的标准化语言能力(水平)测试,大大落后于国际第二语言教学理论和实践的发展。在当时的背景下,北京语言学院(今北京语言大学)的一些有识之士敏锐地抓住时机,提出研制汉语水平考试的设想。在教育部、学校领导的支持下,于1984年底成立“汉语水平考试设计小组”,成员有:刘珣、黄政澄、方立、孙金林、郭树军。设计小组在极其艰苦的条件下,完成了各项研制任务。在1985年第一届国际汉语教学讨论会上,刘珣先生代表研制小组作了题为“汉语水平考试的设计与试测”的报告(见《第一届国际汉语

教学讨论会论文选》，北京语言学院出版社，1986：536～542），引起轰动。1988年6月首次在北京语言学院举行正式考试。1990年HSK通过了教育部组织的专家鉴定，鉴定书认为“HSK的研究成果填补了我国汉语测试的一项空白”。1991年HSK正式推向海外。随着对外汉语教学的发展，原本以汉语预备教育（通常学习时限为一至两年）为主体的局面被打破，短期强化教学（通常为半年左右）和汉语言专业本科学历教育（通常为四年）逐渐发展起来，因此对标准化的汉语水平考试产品提出新的要求。北京语言学院汉语水平考试中心顺应形势的要求，先后完成了HSK（高等）（1993年）和HSK（基础）（1997年）的开发研制，最初研制的汉语水平考试产品改称为HSK（初、中等），三者形成了由低到高的系列考试产品，基本满足了对各类母语非汉语学习者水平测试的需求。

语言测试作为一门交叉型应用学科，以语言学、教育测量学、心理测量学等学科为基础，有着自己的研究对象、概念范畴、研究方法和学科目标。简单地说，这个学科的终极目标是要寻找到对语言能力进行准确测量的理论和方法。而由于人类对自身语言能力的认识还处在极其初级的阶段，这一领域的研究就具有了极大的挑战性。我们强调这一点，是因为社会上对汉语水平考试（HSK）的认识还存在着通俗化或称非科学化的现象，换言之，还有人把她视为科学，认为考试不就是出出题，弄份卷子嘛！因而我们经常会听到这样那样的议论。以题目的难度为例，按照标准化语言测试的基本原理，一个题目的难度，不是由教师、应试者或者其他什么人确定的，而是由该题目在一个较大样本的考生群体中的反应决定的。通俗地说，如果这个群体中的多数答对了，该题目表现为容易，反之为难。此外，对题目的区分度，对试卷的信度、效度等，标准化考试都有严格的要求，因此，评价或衡量一个考试产品的质量，关键要看该产品在上述各方面的数据是否符合科学的要求，只要符合要求，一般来说该产品用来测量语言水平（能力）时就会是准的、稳定的。

汉语水平考试（HSK）多年来一直在科学性上不断追求。围绕着“测

什么”、“怎么测”等基本问题,汉语水平考试中心开展了大量的基础研究。“测什么”讲的是通过语言和语言内容等所表现出的语言能力。要确定能够表现语言能力的语言项目、语言内容,就必须充分考虑各个方面的因素,比如语言材料的题材、体裁、语体、风格等,以及汉字、词汇量、语法点项等。这方面的研究成果主要反映在《汉语水平词汇与汉字等级大纲》(1992)和《汉语水平等级标准与语法等级大纲》(1996)中。“怎么测”讲的是测量技术和测量手段。教育测量学和心理测量学的基本原理和方法是汉语水平考试(HSK)的测试理论和技术基础。无论是题目的设计、试卷结构,还是等级划分、分数解释,都有其依据和标准。在信度、效度、题目难度、区分度,以及公平性等方面都要有理论和数据的有力支持。这些方面的研究成果主要反映在本书系的各篇文章和《中国汉语水平考试(HSK)技术报告》中。可以肯定地说,没有在科学性上的不懈追求和努力,就没有汉语水平考试(HSK)的良好信誉和日益增强的影响力。

汉语水平考试(HSK)作为标准化考试,不仅要做到命题、拼卷、预测、等值等方面的标准化,还要做到考试实施的标准化,这是我们编辑《中国汉语水平考试(HSK)工作规范》的初衷。

本书系的五本书,既各自独立,又互有联系。基本的编辑意图可以表述为:以《语言测试及测量理论研究》表明汉语水平考试(HSK)的基础理论支撑,以《中国汉语水平考试(HSK)研究报告精选》展示汉语水平考试(HSK)的研究成果,以《语言测试专业硕士论文精选》介绍北京语言大学汉语水平考试中心的教学成果,以《中国汉语水平考试(HSK)技术报告》体现汉语水平考试(HSK)的技术实力,以《中国汉语水平考试(HSK)工作规范》说明汉语水平考试(HSK)的操作保证。

汉语水平考试(HSK)作为世界语言测试园地里的一棵小树,已经枝叶茂,与 TOEFL 等语言测试产品一道,服务于世界各国人民语言文化的学习和交流。和任何产品一样,汉语水平考试(HSK)并非十全十美,还

有改进的空间,这已经列入我们的工作计划。而改进同样需要研究,需要建立在科学论证的基础上。社会期待着改进后的汉语水平考试(HSK),我们也将推出改进版的 HSK 的同时,奉献出我们新的研究成果。

中国汉语水平考试(HSK)研究书系

编辑委员会

2005 年 6 月

目 录

1	三类口语考试题型的评分研究	王 佶 旻
28	效度理论的变迁	常晓宇
63	主观评分信度 Longford 方法实验研究	陈菊咏
78	初、中等汉语水平考试试题中的 DIF 检验	黄春霞
103	四种完形填空测试方法的信、效度检验	周 华
129	多种 DIF 检测方法的比较研究	于媛颖
158	对 HSK 三、六级发证标准的验证性研究	徐 静
189	HSK(初、中等)听力部分的分数解释研究	王 洋
215	作为第二语言的汉语写作能力测验方式的实验研究	赵 亮
240	对第二语言阅读能力和写作能力关系的实证分析	马新芳
267	后记	

Contents

- 1 A Scoring Research of Three Different Item Forms
in Oral Testing Wang Jimin
- 28 The Development of Construct Validity Theory Chang Xiaoyu
- 63 The Evaluation and Research of Rater Reliability
with Longford Method Chen Juyong
- 78 DIF Detecting of the Items in Chinese Proficiency Test
(Elementary & Intermediate) Huang Chunxia
- 103 The Comparison of the Reliability and Validity of
Four Kinds of Cloze Tests Zhou Hua
- 129 The Comparison of Several Methods Detecting DIF
..... Yu Yuanying
- 158 A Verifying Research of the Standards of Certificate
for HSK Level 3 and Level 6 Xu jing
- 189 A Research on Scores Explanation of Listening Part
in HSK Elentary & Intermediate Wang Yang
- 215 An Experimental Research of Testing Approaches
to Writing Competence of Chinese as a Second Language
..... Zhao Liang
- 240 An Experimental Research of the Relationship between
Reading Competence and Writing Competence of
Second Language Ma Xinfang
- 267 Postscript

三类口语考试题型的 评分研究

王信旻

■ **内容提要:** 本文通过对三类汉语作为第二语言口语考试题型的评分研究探讨了口语考试评分的客观化问题。我们探讨的客观化评分方法包括 0/1 制评分和在操作性定义口语水平的基础上制定的分项客观化评分。通过实际施测和评分,本研究检验了客观化评分的信度和效度。研究表明,0/1 制评分具有良好的信效度,在其中两类题型的评分中具有较高的可行性。分项客观化评分对保证评分信度有较大贡献,其共时效度也比较好,但是结构效度不是很理想。

■ **关键词:** 汉语作为第二语言 口语考试 评分 信度 效度

Abstract: Using three different Chinese as a second language oral testing methods, this paper makes an inquiry into objective scoring procedures for oral testing. The procedures we explore include right-or-wrong scoring and the so-called "analytic objective scoring" which is based upon our operational definition of oral proficiency. Through experimental study, we find the reliability and validity of right-or-wrong scoring are both fairly good, but the construct validity of "analytic objective scoring" is not so satisfying though its reliability and concurrent validity are high.

Key words: Chinese as a second language, oral test, scoring, reliability, validity

零 引 言

0.1 选题及研究价值

在第二语言口语考试中,研究者们最关心的就是评分问题。现行的口语考试大多采取评分员等级评分的办法,这种评分办法主观性较强,难以保证评分信度,同时也难以对考生之间的水平差别作比较精确的区分。对于同一个言语样本(speech sample),不同评分员之间,同一评分员在不同时间所做出的评判有显著的差异(Lumley & McNamara, 1995)。因而,如何最大限度地排除评分的主观性,提高评分的客观化成为摆在我们面前的难题。

虽然许多学者和研究机构对此问题进行过研究,但这些研究多数是针对评价口语水平的不同等级量表展开的,因此并没有摆脱主观评分的局限,而且大多集中于英语作为第二语言的领域,对汉语作为第二语言口语考试评分问题的研究还不太深入。为此,我们有必要对汉语作为第二语言口语考试的评分问题进行探讨,以期找到客观化的评分方法,提高口语考试评分的科学性。

0.2 解题

0.2.1 三类题型的选择

根据研究的需要和测验实施的方便,我们选择了三种题型,分别是:

- A. 问答(Question and Answer)、B. 重复句子(Sentence Repetition)和
- C. 口头报告(Oral Report),以下分别简称为A题型、B题型和C题型。

我们设计的A题型是一种快速问答,考生听到问题后必须在5~10秒内回答。这种题型具有较好的表面效度(face validity)*,操作时指导语简洁,考生易于理解,而且可以独立评分。

* 表面效度指测验在表面上使被试直觉感到的有效性程度。如果一个测验使被试从表面上看来与测验的目的无关,被试就会对测验缺乏信任从而就会减弱他努力完成测验的动机和积极性,这样的测验就被认为缺乏表面效度。

B 题型由互相独立的 40 个句子组成,每句字数为 5~14 字*。从心理语言学的观点看,重复句子是一种“组块”**(chunk)过程。语言水平高的人能用短时记忆储存和重组所听到的信息(Clark & Swinton, 1979),因此这一题型很值得心理语言学界和心理测量学界研究。

C 题型是在非面试环境下诱导考生说出较长的一段话,这种题型具有较高的表面效度,能考查考生的综合表达能力。TSE、HSK(高等)口语和我国英语专业四级口语考试等都包含有这种题型。需要说明的是,本研究设计的口头报告是根据指定的话题说话。

0.2.2 对客观化评分的理解

Bachman(1990)从评分的角度区分了客观考试和主观考试,认为客观考试是指考生反应的正确性完全由既定标准决定而不需要其他判断;主观考试是指评分员必须对考生反应的正确性作出判断,而这种判断是基于评分员对评分标准的主观解释上的。由此我们可以发现,客观考试和主观考试的区别主要在于对评分标准的解释上,前者的解释是惟一确定的,后者的解释则随评分员理解的不同而有不同。因而,要保证评分的客观性就必须最大限度地保证对评分标准解释的确定性,减少评分员主观判断的成分。本文所指的客观化就是评分标准的客观化,即提高对评分标准解释的确定性。

0.3 研究目的

本文希望通过对三类汉语作为第二语言口语考试题型的评分研究来探讨口语考试评分的客观化问题。我们欲探讨的问题有:

(1) 0/1 制的客观化评分:即在口语考试中是否可以采用 0/1 制评分,0/1 制评分的口语考试信、效度如何。

(2) 非 0/1 制的客观化评分:即在不适用 0/1 制评分的情况下如何

* 我们选择 5 作为句子字数的起点是因为短时记忆的容量为 7 ± 2 ,即一般为 7,并可在 5~9 之间波动。(Miller, 1956, 转引自王甦、汪安圣, 1992)

** 组块是指把小单位联合成较大单位的信息加工。

找到客观化的评分标准,采用客观化标准评分的口语考试信、效度如何。

(3) 评分方法和题型的结合问题:讨论在本文涉及的评分方法中哪一种最适合 A 题型,哪一种最适合 B 题型,哪一种最适合 C 题型。

一 客观化评分方法的探索

——理论基础及操作性定义

制定一种有效的评分方法需要解决两方面的问题:第一是定义,即决定什么意味着懂得了一门语言;第二是方法,即寻找正确的测量过程(Shohamy, 1998, 见 Byrnes, 1998)。Bachman & Palmer (1996)更加具体地指出了建立一种评分方法的三步曲:(1)对所要测量的结构(construct)进行理论定义,(2)对结构进行操作性定义,(3)建立对被试的回答进行量化的方法。下面我们将依据这三个步骤来探讨口语考试的客观化评分方法。

1.1 口语水平的理论定义

Brown(1996)在模式(mode)和渠道(channel)两个层面上区分了口语、听力、阅读和写作四种语言技能。Brown认为模式有两种,一种是接受性的(receptive),另一种是产出性的(productive);渠道也有两种,一种是书面的(written),另一种是口语的(oral)。四种语言技能在模式和渠道上的表现是不同的,可以用图表直观地表示如下:

		渠 道	
		书面	口语
模 式	接受性	阅读	听力
	产出性	写作	口语

图 1

从图表中我们可以看到,口语渠道和书面语渠道都有接受性和产出性两种模式。正是通过这两种模式的相互作用,一定渠道的交际任务才得以完成。因此从口语渠道来讲,它是听和说相互作用的结果,说话人依靠听的能力来接受信息,依靠说的能力来产出信息。

口语考试关心的是对口语水平的测量,因此我们必须明确界定口语语言水平(oral language proficiency)。对语言水平(language proficiency)大致有两种理解,其一是把它等同于语言能力(language competence),指内化了的语言知识(Ellis,1985);其二是把它看做学习者使用第二语言的能力(Oller,1979,Bachman,1990,Ellis,1994)。我们在这里把语言水平定义在第二种理解的基础上,即学习者使用第二语言的能力。

根据对口语交际渠道和语言水平的理解,我们把第二语言口语水平定义为学习者在听说模式的口语渠道中使用第二语言口语的能力。

1.2 口语水平的操作性定义

在口语考试中,对口语水平的操作性定义集中表现在不同的水平等级量表上。这些量表分为两种类型,反映了对口语水平的两种不同的操作性定义。第一种为总体水平等级量表(holistic proficiency rating scale),它把口语水平操作性地定义为学习者的总体口语表现,如FSI总体水平等级量表、ILR量表和ACTFL量表。第二种为分项水平等级量表(analytic proficiency rating scale),它把口语水平操作性地定义为学习者在几个方面的口语表现,如FSI分项水平等级量表、TSE在1995年以前使用的分项水平等级量表等。

从总体口语表现来推断其口语水平还是从口语表现的几个方面来推断其口语水平,这是两种操作性定义的主要分歧。我们认为,在评价口语水平时可以将其分为几个方面来考虑。这是因为从口语习得的过程来看,发音是最基本的,语法的准确性(accuracy)和可理解性(comprehensibility)要在一定的发音基础上才能实现,而流利性则要在前两者的基础上才能达到。这说明发音、语法和流利性这几方面的发展是不平衡的,因此在评价时应该区别对待。从评价口语水平的角度来看,人们在评价口