HANYU KOUYU CESHI
PINGFENYUAN PINGJIA YANJIU

汉语口语测试评分员评价研究

黄霆玮◎著

中国书籍出版社
China Book Press

HANYU KOUYU CESHI
PINGFENYUAN PINGJIA YANJIU

汉语口语测试评分员评价研究

黄霆玮◎著

中国书籍出版社
China Book Press

图书在版编目（CIP）数据

汉语口语测试评分员评价研究 / 黄霆玮著. —北京：
中国书籍出版社，2020.1
ISBN 978-7-5068-7553-0

Ⅰ. ①汉… Ⅱ. ①黄… Ⅲ. ①汉语—口语—测试—研究 Ⅳ. ①H193.2

中国版本图书馆CIP数据核字（2019）第270092号

汉语口语测试评分员评价研究

黄霆玮 著

责任编辑 王志刚
责任印制 孙马飞 马 芝
版式设计 中尚图
出版发行 中国书籍出版社
地　　址 北京市丰台区三路居路97号（邮编：100073）
电　　话 （010）52257143（总编室）（010）52257140（发行部）
电子邮箱 eo@chinabp.com.cn
经　　销 全国新华书店
印　　刷 河北盛世彩捷印刷有限公司
开　　本 710毫米×1000毫米 1/16
字　　数 210千字
印　　张 14.5
版　　次 2020年1月第1版 2020年1月第1次印刷
书　　号 ISBN 978-7-5068-7553-0
定　　价 58.00元

序 言

普通话水平测试是世界上规模最大的口语测试。在近40年的发展历程中，普通话水平测试推动了普通话推广工作的深入发展，同时也推动了口语测试的发展和研究。针对测试实践中出现的各种问题，许多研究者从不同的角度做了论述。在此过程中，普通话测试逐渐发展成为应用语言学的一个分支，在不断发展壮大。2006年，社科院成立了普通话测试方向的博士点，我开始招收普通话测试方向的研究生。

2008年，霆玮成功通过了社科院的博士生入学考试，成为我普通话测试方向的第二个博士研究生。博士点建立伊始，可以说是筚路蓝缕，不仅没有现成的专业书籍，甚至也没有完整的理论体系。我带领我的博士生们一边学习一边探索，出版了《普通话水平测试概论》一书。这本书搭建了普通话测试的基本理论框架，提出普通话测试有测试员、测试对象、测试依据和测试手段等四个要素。

当时在讨论过程中，霆玮就对普通话测试员感兴趣，她认为测试员在口语测试中的作用是决定性的，测试的其他几个要素都要通过测试员起作用。在积累了相当的资料后，她选定测试员作为自己的研究方向。我们经过数次讨论，将研究的切入点定在对测试员的评价上，希望能推动测试员的管理工作更加规范化、科学化，对各级普通话测试站的测试员培训工作有所助益。

霆玮硕士期间读的专业也是语言测试，有较好的测试学理论基础，

也做过一些测试的实践工作，对语言测试的理论和实践都有系统的了解。她想把自己的专业特长与普通话水平测试的实际结合起来，并把研究的范围扩展到整个汉语口语测试领域，最终选定了“口语测试评分员评价”这个题目。完成这样一项任务，对于一个博士生来说并不轻松，这中间她克服了很多困难，终于如期完成了论文。

本书以霆玮的博士论文为基础，经过几年的沉淀积累，现在呈现在读者面前。在普通话水平测试研究不断深入的过程中，本书在以下几个方面推动了口语测试的研究与发展：

（1）本书将现代测量学理论运用到普通话测试中，使用项目反应理论（IRT）刻画测试员的评分误差。这是一个崭新的视角，将测试员平时不可见的评分特征展示了出来，为评价测试员的工作提供了科学依据。

（2）本书采用了问卷调查、多元统计等多种实证研究方法，为评分员改进评分工作提供了具体可行的方法，在此基础上，为评分员的评价提供了客观的依据，为评分员管理机构的决策提供了科学依据。

（3）当前语言测试领域多研究评分员评分的信度、效度等方面，本书以对评分员评价为研究对象，拓展了评分员研究的领域，丰富了测试学研究的内容。。

为人师者，如同为人父母者一样，无不期望学生成龙成凤，在自己的专业领域内有所建树。霆玮这些年一直坚持做语言测试领域的研究，这本书是一个小小的里程碑，希望她在治学的道路上能以此为新的起点，脚踏实地，取得新的更多的成绩！

姚喜双

2019年12月

目 录

图表目录

第1章　绪论

1.1　研究缘起

1.1.1　口语测试的发展

在语言测试中，口语测试是一种常见的考试类型，是测量应试人口语能力最直接的一种手段。20世纪末期，Bachmam提出了著名的“语言交际能力说”。这种语言能力观认为语言能力不仅包括对语言系统知识的掌握，还包括对句子之外语言交际环境的掌握[①]。基于“语言交际能力说”的语言测试体系强调测试的“真实性”和“交际性”。在这种背景下，口语测试因其符合真实性和交际性的特点，日益受到重视。

口语测试是一种主观测试。与客观测试相比较，口语测试命题简单，评分却比较困难。口语测试在真实的交际环境中进行，评分误差的来源比较多。如何控制口语评分的误差，保证口语考试的信度是主观性考试中的一个重要课题。

主观考试评分中的误差主要来源于测试任务、评分标准、评分量表和评分员等方面。测试任务、评分标准和评分量表等都是测验的开发者制订的，

① 陈菁：《从Bachman交际法语言测试理论模式看口译测试中的重要因素》，《中国翻译》，2002年第1期，第52页。

处于测验开发者可控制范围之内，测验开发者可以不断修改、逐步完善。而评分员是测验开发者无法把握的一个误差来源，评分员的表现可能受到各种因素的影响，是动态的、不断变化的。评分员评分是一个根据既定的评分标准和评分量表，给应试人口语能力赋值的过程。评分标准和评分量表要通过评分员才能作用于应试人。评分标准和评分量表被评分员理解、内化，最后才应用于被试。所以，评分员如何评分直接关系到口语测试的信度和效度，评分员的评分质量是测验开发者的设计思路能否实现的关键。很多研究显示，不同评分员评分的过程差异很大[①]。评分员在理解、内化评分标准时发生了什么？产生了哪些差异？如何描写这种差异？不同的评分员差异反映的本质是什么？

进而我们要讨论：这些评分员差异对评分质量有哪些影响？什么样的评分员的评分质量较高？什么样的评分员评分质量较差？我们应该如何评价一个口语测试的评分员？这就是本书要讨论的问题。

1.1.2 评分员评价体系研究现状

在主观测试领域中，对评分员的研究一直是一个热点。这些研究的角度不同，有关于评分员的评分方法的，有关于评分员的评分信度的，还有关于如何培养评分员的，但其中有关汉语口语测试评分员评价的研究不多。在我们搜集到的文献中，仅有三篇是专门研究普通话评分员考核的，与我们要探讨的评分员评价体系研究比较接近。

① Eckes, T.2008. Rater types in writing performance assessments: A classification approach to rater variability, *Language Testing*, 25(2).

毛立群（2003）[①]主要探讨了普通话水平测试员考核体系的建立。文章首先从以下三个方面归纳了测试员队伍的现状：业务素质、职业道德和科研进修。在此基础上结合浙江省普通话水平测试员管理的经验，提出了建立普通话水平测试员考核体系的设想，包括以下四点：规范选拔程序，保证选送人员的质量；点面结合，使业务素质的考核尽量做到量化；工作量考核能客观反映出测试员的热情和态度；强调科研进修，确立后续培训制度。这篇文章从普通话水平测试实践管理出发，较全面地论述了普通话水平测试员考核体系的内涵。美中不足的是，这篇文章比较宏观，没有往深处挖掘考核评分员的具体指标以及考核评价对评分员的反馈效果。

钱华（2004）[②]的研究，是迄今为止有关普通话水平测试员考核体系的研究中较为全面的一篇。文章首先从测试实践出发，总结归纳了测试员考核中存在的问题，在此基础上提出构建测试员考核体系的意义和原则，其次提出了测试员综合指标体系的内容与基本框架，最后论述了考核工作的组织实施以及考核结果的运用。这篇文章的考核指标体系涉及四大方面：思想素质结构、业务素质结构、身心素质结构和绩效结构。这四个方面作为考核体系的一级指标，每个一级指标又具体细化为若干二级指标，最后呈现为26个三级指标。这些指标设定得非常全面，包含了《国家语言文字工作委员会关于普通话水平测试管理工作的若干规定（试行）》第十一条规定的普通话水平测试评分员的考核内容：工作态度、测试能力、测试工作量、遵守工作纪律情况等。同时，此研究还提出了考核的具体实施步骤。这篇研究从普通话水平测试员的测试实践出发，具有很强的参考价值，但是理论的部分还有待加强。

---

① 毛立群：《试论普通话水平测试员管理考核体系的建立》，苏培成编，《中国语文现代化学会2003年年度会议论文集》，语文出版社，2003年版，第251—258页。

② 钱华：《普通话水平测试员综合考核指标体系构建研究》，国家语言文字工作委员会普通话培训测试中心编《第二届全国普通话水平测试学术研讨会论文集》，2004年版，第81—91页。

在汉语水平考试（HSK）高等口语测试的相关研究中，专门对口语测试评分员展开的研究不多。有些研究的成果可供参考，例如：关于评分误差控制、评分员培训的研究，但还没有见到专门关于口语测试评分员评价的研究。

1.2 研究思路

1.2.1 研究目的和内容

本文以语言测试学、人力资源评价理论为指导，采取理论与实证相结合的方法，通过研究旨在揭示评分员评价的本质，提出构建汉语口语测试评分员评价体系的理论依据，确立汉语口语测试评分员评价体系的指标，设计评价方案，从理论与实践两个层面提出解决汉语口语测试评分员评价的理论体系和实施方法。

从选题视角引出口语测试、语言能力、评分员、评分员评价等基本概念，对这些口语测试中的基本概念及它们的特征做详细论述，在此基础上完成构建包括“素质、能力、绩效”为一级评价指标的评分员评价体系。对“素质”、“能力”和“绩效”的评价分别通过“考核”、“考试”和“考绩”的方式进行。三种评价体系中，对“绩效”的评价是最重要的，其他两种处于辅助地位。本文的主要研究内容包括：

（1）构建汉语口语测试评分员评价模式。阐述了建立汉语口语测试评分员评价体系的理论基础，包括建立评价体系的目的、作用、原则等。一个完整的汉语口语测试评分员评价体系包含三个部分：“素质”评价体系、“绩效”评价体系和“能力”评价体系。在这三个方面中，“素质”和“能力”主要是用来衡量评分员的内在价值，“绩效”主要是衡量评分员的外在价值，也就是

评分员创造的价值。内在价值能够转化为外在价值，所以在三个一级指标中，“绩效”指标是最直接和最主要的。我们认为，这三个方面较全面地代表了评分员的日常工作表现，是一个具有实际应用价值的理论框架。

（2）构建评分员“绩效”评价模式。对评分员“绩效”的评价主要反映在对评分员评分质量的评价上。评分员的任务很多，特别是普通话水平测试（PSC）的评分员还有推广普通话等其他任务。不同口语测试中，评分员承担的任务不同，但其主要任务是为应试人评分。评分质量的高低关系着口语测试的信度和效度。本文为了量化评分员的评分质量，构建了以严厉度、一致性为指标的“绩效”评价体系。这个体系在理论上能够反映评分员评分结果和应试人能力的差别，可以用来评价评分员的评分质量。

（3）应用评分员“绩效”评价模式进行实证研究。本部分将使用现代测量理论尝试量化评分员的评分质量，为评价评分员提供测量学方面的理论支持。本部分的另一个贡献是对量化结果进行了有效性检验，检验结果显示严厉度和一致性作为评价指标可以反映评分员的评分质量。

1.2.2 研究方法

理论与实证研究相结合的研究方法是本文研究最基本的研究方法。具体来讲，本文使用的主要研究方法有：

（1）文献法。为完成本项研究，我们搜集了数百篇有关汉语口语测试、评价体系的学术论文，穷尽性地收集了关于普通话水平测试（PSC）和汉语水平考试（HSK）的学术论文，其中包括数十篇硕博士论文。除此之外，笔者还认真研读了语言测试方面的中外文专著。通过阅读文献，掌握了进行评分员评价的理论和方法，为完成论文打下了良好的基础。

（2）分析法。在占有大量文献资料的基础上，“去粗取精、去伪存真、由