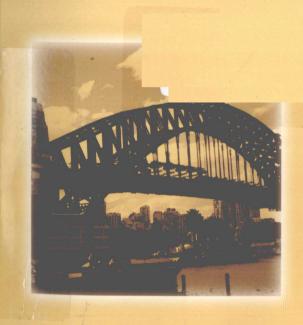# 中国大学生英语口语分析性评估体系的构建与效验

Construction of and Validation Study on Analytic Rating Scales for Oral English Proficiency of Chinese Tertiary Learners

● 刘 芹 著

中国大学生英语口语分析性评估体系的构建与效验

Construction of and Validation Study on Analytic Rating Scale for Oral English Proficiency of Chinese Tertiary Learners

复旦大学出版社

中国大学生英语口语分析性评估体系的构建与效验
刘　芹　著

内容提要

本书作者以Bachman的交际语言能力（CLA）模式、Cohen的语用口试等级评分表、Nunn的小组讨论等级评分表及高校英语教学大纲为基础，结合大规模问卷调查结果构建了中国大学生口语评估体系。作者在将其运用于来自六类院校180名学生的口语评估后，对其进行了全面的效度研究，同时深入探讨了中国大学生英语口语的普遍特征。作者提出的口语评估体系具备科学性和可行性，对口语测试评分标准的改进和完善具有借鉴意义，对提高口语测试的总体效度有着很高的理论意义和应用价值。本书的读者对象为从事英语口语教学评估的研究人员和专业学生，也可供高校英语教师参考使用。

序

　　口语能力是语言能力的重要组成部分,其测评历来是语言测试研究者和语言教学工作者所关注及研究的重点之一。口语测评大多是一种基于被试者表现的考试形式(performance-based testing),即被试者根据考试要求,完成一个或数个口语任务,然后由考官根据考生表现决定其口语能力等级。在这个过程中有几个因素会影响考生表现:首先是他的口语能力,其次是口试形式、评分标准以及考官个人因素。其中一个基础且关键的环节是评分标准。因而,如何制定一个能客观、公正、全面地衡量口语能力的评分标准是口语测试人员迫切需要解决的一个课题。

　　本书在口试评分标准制定方面做了卓有成效的探索。作者以分析性评分模式为切入点,以多种理论模式为依据,以现行英语教学大纲为基准,在大规模调研的基础上,提出了分析性口语评估体系(OARS)。

　　本书具有以下特点。文献部分的综述全面、详尽,涵盖了本领域的最新研究成果。作者从不同角度阐述了相关理论模式,包括语言能力模式及口语评分理论依据。阐述条理清晰、概括性强、层次分明,显示了作者扎实的理论功底。口语评估体系的第一个特点是测量维度全面:既包括语言能力又涵盖策略能力。同时,语言能力的衡量能兼顾到口语能力的外显特点(如语音语调等)。策略能力的判断也基于口语交际的特征之上,如包括体势语、声调、音量等。因此,该评估体系能较为客观、公正地评价口语能力。口语评估体系的另一特点是兼顾英语教学要求,每个小项都有相应的等级描述。这有利于该体系的使用者(教师、学生)充分利用口语测试的信息反馈,促进教与学。另外,研究设计全面,研究方法科学,对口语评估体系的效度验证

过程严谨,结论合理,论证结果令人信服,显示了作者良好的科研素质和能力。

　　总而言之,本书是一项高质量的研究成果。作者提出的口语评估体系具备科学性和可行性,对口语测试评分标准的改进和完善具有借鉴意义,对提高口语测试的总体效度有着很高的理论价值和应用价值。

<div align="right">

邹申

上海外国语大学教授、博导

教育部高等学校外语专业教学指导委员会委员

英语专业教学分指导委员会副主任委员

2009 年 11 月

</div>

前　言

中国大学师生历来十分重视英语口语能力的培养。英语专业和大学英语教学大纲对每一级别学生均提出具体的英语口语能力要求。然而,大学生是否达到了大纲的要求呢? 蔡基刚(2002)转引四、六级考试委员会对于 1999 年 1 月至 2001 年 5 月的六次口试(CET-SET)情况的总结报告指出,在累计32107位考生中,获得 B 以上的考生是18550人,即只有 57.8% 的考生能用英语就熟悉的题材进行口头交际。"如果按照每次 200 万四、六级考试考生来算,其百分比极小,可见大学生英语口头交际能力很差。"另一方面,文秋芳、吴彩霞和 Lydia So(1999)认为"除在讲话语速方面我国英语专业二年级学生已超过大纲要求外,其他各项都难以说明是否达到了大纲要求。"文秋芳、赵学熙和王文宇(2001)进一步指出 2000 年专业英语四级口试(TEM4-Oral)中存在语言准确性偏低,语言流利度不够,讲话内容没有新意,缺乏思维的深度,和对话中交际原则掌握不好等问题。

从两项口试的情况来看,中国大学生的英语口语能力并不令人满意。而且,参加大学英语口试的学生是笔试成绩优秀的考生,参加英语专业口试的学生每年只有几千人。相当数量的学生并未参加这两类口试,也无从衡量自己的英语口语能力,更无法与其他院校或专业的学生作横向比较以发现自身的强项和弱点。如何全面、详尽地衡量学生的英语口语能力,为中国高校英语口语教学提供较为详尽的反馈信息,以切实提高学生用英语进行口头交际的能力,这是迫在眉睫的一个研究课题。

本书旨在针对上述问题提供一套分析性口语评估体系(OARS)。该体系以 Bachman 的交际语言能力(CLA)模式、Cohen 的语用口试等级评分表、Nunn 的小组讨论等级评分表、

英语专业及非英语专业英语教学大纲、以及对 182 位教师和 1 139位学生的大规模问卷调查结果为基础设计而成。就测量难度而言,它主要分为语言能力和策略能力两大方面。语言能力方面评估两个特质:组织能力(organizational competence)和语用能力(pragmatic competence)。前者考察言内组织能力,即语音(pronunciation)水平——受试的语音语调是否准确及达到本族语者水平;语法(grammar)水平——受试的话语在词汇和句法上是否正确;语篇连贯性(textual coherence)——受试是否能运用恰当的连接方法将语篇连成一体。后者考查言外语用能力(illocutionary pragmatic competence)——受试能否清晰表达意思并快速对他人话语作出合理反应;社会语用能力(sociolinguistic pragmatic competence)——受试的话语是否符合语言使用环境(如方言、语域和俗语)。策略能力反映在对灵活交流(flexible interaction)和非言语交际(nonverbal communication)这两个特质的评估上。前者测量受试使交流平稳进行的能力。为达到这一目的,他/她必须展示话轮转换能力(例如取得话轮、保持话轮、传递话轮等)。后者从两方面进行评价:体势语(body language)——受试在讲话时是否能合理使用目光接触、手势、面部表情等体势语更好地传递信息;和副语言(paralanguage)——受试是否能运用非言语声响(如声调、音量、重音等)帮助自己表达意思。在设计形式上,OARS 共有八个小项,每项分五个等级(5 分为最高,1 分为最低),总分 40分。每个小项配备详细的等级描述语。

　　语言评估是与实用性紧密相连的一门学科。理论上设计完好的评估体系在实际应用时往往会发生问题。因此为检验 OARS 的信度、效度和实用性,笔者对其进行了如下两阶段的效度研究。

　　第一阶段是小范围效度研究。它主要涉及该评估体系在应用上可能存在的问题及其信度和效度。上海理工大学(USST)的 30 名英语专业二年级学生参加了本次研究的试测工作。他

们被随机分成三位一组,每组进行八分钟的讨论。试测全程录像后交由两位评分员按照 OARS 的等级描述语打细评分(即分析分),另由六位教师根据 OARS 的框架打粗评分(即印象分)。然后求出细评分的平均分作为整项研究的核心。考生在试测同期参加的英语专业四级考试(TEM4)和口语课程大考成绩以及两年后参加的英语专业八级考试(TEM8)成绩亦记录在案,作为外部效标。经过信度检验、正态分布检验及相关检验得出如下结论。

首先,OARS 的信度系数为 0.7902,符合口语评估标准。第二,绝大多数部分－整体相关系数超过 0.7,而部分－部分相关系数低于 0.6,显示合理的构想效度。第三,细评分平均分与粗评分平均分及口语课程大考成绩的相关系数分别达到 0.831 和 0.653,表明较高的外部效度。另外,参加该项研究的师生对 OARS 持肯定态度,他们普遍认为 OARS 可以全面详尽地评估中国大学生的英语口语能力。而且,小项及等级描述语都通俗易懂。从定量和定性分析两方面来看,OARS 是一套高信度、高效度且颇具实用价值的英语口语评估体系。

第二阶段的大规模研究探讨 OARS 的延伸应用问题以及中国大学生英语口语能力的特征。180 位刚通过英语专业考试(TEM)四级或大学英语考试(CET)六级的大学生参加了此项研究的试测工作。他们分层抽样于以下六类学生(每类 30 名):一本院校英语专业学生、一本院校文科学生、一本院校理科学生、二本院校英语专业学生、二本院校文科学生和二本院校理科学生。然后把他们随机分成三人一组,每组进行八分钟的讨论。试测全程录像后交由两位评分员按照 OARS 打分析分,并求出平均分。结果显示较高的信度(每类的信度系数分别为 0.7900、0.7885、0.7693、0.7902、0.7853 和 0.7754)和构想效度(大多数部分-整体相关系数达到 0.7 的理论要求)。结论如下:OARS 可以高信度、高效度地应用于不同类型学生的英语口语能力评估。

　　当将180位考生作为整体样本时,OARS的信度系数和构想效度更为理想,表明在考生数量大的情况下,该评估体系具有更强的信度和效度。对全体考生的试测情况反映出他们在英语口语能力方面的一些共性的问题。第一,未达到高层次语音要求。OARS在"语音"上的评估包括单音素的发音和合适的发音技巧与语调。结果显示,相当一部分考生虽然在单音素上不存在问题,但他们发音生硬,尚不能掌握同化、略读、连读、重读、失爆等高层次的语音要求,更谈不上如本族语者的发音了。第二,尚未掌握交际原则。OARS从语用能力(言外语用能力和社会语用能力)和灵活交流两方面对言语表述能力作具体的衡量。结果显示三分之二左右的考生能够清晰表达意思并快速对他人话语作出合理反应,而且用语比较符合语言的使用环境。然而,在灵活交流能力方面还是不容乐观。近一半的考生缺乏交际手段,他们不能找到合适的切入点参与讨论,有的胡乱打断他人发言,有的一有发言机会就喋喋不休,也有的始终被动等待发言机会。第三,缺乏非言语交际意识。诸多语言学家和教学评估专家纷纷指出,非言语交际能力是交际能力中不可或缺的一部分(如 Harrison 1965, Morlan & Tuttle 1976, Canale 1983, Canale & Swain 1980, Ross 1986, Nunn 2000, Taylor 2006 等)。这是西方人交谈时非常普遍的特征,然而一半左右参加本项试测的考生缺乏这一能力。他们在发表自己意见时声调平淡、没有起伏;与他人讨论中没有直接目光接触,面部表情呆板,不会用手势或手势不灵活。以上三个问题在中国大学生中普遍存在,亟待在英语口语教学中引起重视。

　　进一步的对比研究 t 检验和多元方差分析显示:一本院校学生在"语法"、"语篇连贯性"、"言外语用能力"、"灵活交流"、"体势语"和"副语言"等特质上的平均得分显著高于二本院校学生;英语专业学生在所有特质上的平均得分均显著高于非英语专业学生;文科专业学生和理科专业学生的得分没有显著性差别;在一本院校中,英语专业学生仅在"语法"特质上强于非

英语专业学生；在二本院校中，英语专业学生在"语篇连贯性"、"灵活交流"、"体势语"和"副语言"等特质上强于非英语专业学生。

　　本书作者以博士论文为基础，2006 年在美国加州大学洛杉矶分校访学期间对若干章节进行了改写，回国后又结合最新语言评估理论和实证研究成果进行了修订。可以说本书是笔者近年来从事口语评估研究的阶段性总结。

　　在全书撰写期间，笔者得到众多前辈和同行的指点和帮助。首先要特别感谢恩师上海外国语大学邹申教授。邹教授虽然日理万机，但在本书选题、研究设计到撰写过程中她都给予了笔者无微不至的关怀和指导，为本书的成文打下了扎实的基础。北京外国语大学的文秋芳教授百忙拨冗，与笔者讨论了口语评估体系的制订原则。美国加州大学洛杉矶分校的 Lyle F. Bachman 教授在笔者访学期间对实验数据的统计分析提供了帮助。上海交通大学的金艳教授和广州外语外贸大学的曾用强教授对本书的最后定稿提出了宝贵建议。笔者在上海外国语大学就读博士期间，得到了诸多教授和同学的帮助。复旦大学出版社为本书的出版提供了热忱的支持。在此一并致谢。

　　上海外国语大学、复旦大学、上海财经大学、华东理工大学、同济大学、上海师范大学和上海理工大学的数百名教师和一千余名学生在本研究进行过程中参加了问卷调查、口语试测和评分，笔者对他们的无私奉献深表感激。

　　笔者的家人在这几年来一直理解和支持我的研究工作，谨表示最崇高的敬意。

　　尽管本书经过了若干次的修改，但笔者才疏学浅，书中难免存在不妥和疏漏之处，恳请各位专家、同行批评指正。

<div style="text-align: right">

刘芹

2009 年 10 月于上海

</div>

Preface

College teachers and students in China have always attached great importance to the cultivation of oral English proficiency. The teaching syllabuses of English majors and non-English majors, therefore, incorporate specific oral English proficiency requirements for students at each level. However, have the students met the requirements set by the syllabuses? Cai Jigang (2002), quoting from reports of CET Committee, points out that from January 1999 to May 2001, 18,550 students got a B in the six oral tests for non-English majors (CET-SET) out of a total number of 32,107, accounting for only 57.8%. He further argues that "Compared with 2 million test takers who take part in each CET written test during this period, this passing rate is extremely low, indicating poor oral English communicative ability among college students." Meanwhile, Wen Qiufang, Wu Caixia and Lydia So (1999) conclude that "Second-year English majors have not definitely reached the requirements for oral English proficiency set by the teaching syllabus except for speaking speed" (34). Taking the results of TEM4-Oral 2000 as an example, Wen Qiufang, Zhao Xuexi and Wang Wenyu (2001) further argue that students have got such common problems as low accuracy, insufficient fluency, a lack of novelty and profundity, and unsatisfactory mastery of communicative rules while holding dialogues and discussions.

It can be concluded from the results of the above-mentioned two national oral English tests that the oral English proficiency of Chinese college students is far from satisfactory. Moreover, only students who have got excellent marks in the CET written test are eligible to take CET-SET and only a few thousand actually sit for TEM-Oral each year. A large number of students have got no experience of taking them. Thus they have got no means to evaluate their oral English proficiency, let alone comparing with students from other majors or universities in order to figure out their strengths and weaknesses. Therefore, it is a research project in urgent need to evaluate students' oral English proficiency in a complete and detailed way so that valuable information can be provided concerning oral English instruction in Chinese colleges and universities to help enhance oral English proficiency of students.

The study presented in this book aims at establishing a model of oral analytic rating scales (OARS) as an attempt to solve this problem. OARS is designed on the basis of Bachman's Communicative Language Ability (CLA) model, Cohen's Rating Scale for Pragmatic Speaking, Nunn's Rating Scales for Small Group Interaction, teaching syllabuses of English majors and non-English majors, as well as questionnaire survey results of 182 English teachers and 1, 139 students. It covers two main aspects of oral proficiency — language competence and strategic competence. The former incorporates organizational competence (pronunciation, grammar and textual coherence, to be exact) and pragmatic competence (illocutionary as well as sociolinguistic competence) while the latter evaluates flexible interaction and nonverbal communication (namely body language and paralanguage). *Pronunciation* tests whether the candidate can

provide utterance with accurate pronunciation and intonation. *Grammar* assesses the candidate's capability of *producing* grammatically correct utterances (in terms of both vocabulary and syntax). *Textual coherence* assesses whether the candidate can use cohesive devices appropriately so as to form coherent utterances. *Illocutionary pragmatic competence* assesses whether the candidate can express ideas clearly, understand others' utterances and give timely and appropriate rejoinders. *Sociolinguistic pragmatic competence* examines whether the candidate's utterance is related to features of the language use context, such as appropriate uses of dialects, registers, and idiomatic expressions. *Flexible interaction* assesses whether the candidate has the ability to keep an interaction going smoothly. In doing so, he/she may have to exhibit the ability for taking turns (such as holding the floor, interrupting politely, and helping other participants to join in) and negotiating (such as taking initiatives and appreciating other participants' contributions). *Body language* includes eye contact, gestures and facial expressions while *paralanguage* refers to the vocalizations that are not words but that may convey meaning or add to the meaning of words (such as tone, volume and stress). Each of the eight traits (hereafter referred to as "Pron", "Gram", "Text", "Illoc", "Socio", "FlexI", "Body" and "Para" for convenience) is provided with a scale from 5 as the highest to 1 as the lowest, together with detailed band descriptors.

Language assessment is an endeavor closely linked to practicality. Problems do occur when a theoretically well-defined assessment model is put into practice. Thus, OARS has undergone the following two stages of validation study to prove its reliability, validity and utility.

The pilot study is mainly concerned with the feasibility of this

model, possible problems with regard to its application and its reliability and validity. It was conducted on 30 randomly selected sophomores majoring in English in University of Shanghai for Science & Technology (USST). They were randomly arranged into groups of three and each group was required to hold an 8-minute discussion. The picture recording of their performance was evaluated by two raters who did detailed ratings according to the band descriptors of OARS and six teachers who did rough ratings according to its framework. Average scores of the first kind of rating were taken as the core component of the study. The candidates' scores on TEM4 and oral English final examination, taken at roughly the same time of the experimental test, as well as their scores on TEM8 taken two years later, served as outside measures. Results were gained from tests of reliability, normality and correlation, shown as follows.

Firstly, OARS exhibits an overall reliability coefficient alpha of 0.7902, quite desirable for oral proficiency assessment. Secondly, most sub-total correlation coefficients are above 0.7 while most sub-sub correlation coefficients are below 0.6, indicating satisfactory construct validity. Thirdly, the correlation coefficient of the detailed ratings and the rough ratings reach 0.831 whereas that of the detailed ratings and the scores on the oral English final exam reach 0.653, signifying acceptable external validity. In addition, both the candidates and the teachers who conducted ratings expressed their approval of OARS for it can evaluate the oral English proficiency of Chinese college students in a detailed and comprehensive way. They also found it easy to follow the scales and band descriptors of OARS. So qualitatively and quantitatively, OARS establishes itself as a model of reliable, valid and feasible oral English rating scales.

The subsequent field study mainly focuses on the further applicability of OARS, its reliability and validity when applied in large-scale oral English assessment, and characteristics of Chinese college students with regard to their oral English proficiency. 180 candidates who have just passed TEM4 or CET6 took part in this study. They were selected from six levels of college students, namely English majors of key universities, Arts majors of key universities, Science majors of key universities, English majors of local universities, Arts majors of local universities, and Science majors of local universities. As in the pilot study, they were randomly grouped into three and required to hold an 8-minute discussion. All their performance was picture recorded and rated by two raters by means of OARS. Results display satisfactory reliability coefficient alphas (0.7900, 0.7885, 0.7693, 0.7902, 0.7853 and 0.7754 respectively) as well as construct validity (with most of the sub-total correlation coefficients reaching the theoretical demand of 0.7). It is concluded that OARS can be applied to different kinds of Chinese college students in the assessment of oral proficiency with acceptable reliability and validity.

When applied to 180 candidates as an entire sample, OARS exhibits even more satisfactory reliability coefficient and construct validity, indicating its higher reliability and validity while applied to large number of candidates. Common problems of these candidates concerning oral English proficiency are reflected in the following. The first is that candidates have not acquired advanced pronunciation skills. The trait "Pron" in OARS is meant to assess accurate pronunciation of vowels and consonants as well as appropriate pronunciation techniques and intonation. Results show that a large number of candidates have reached the basic

requirement of accurate pronunciation of phonemes, but their pronunciation and intonation have remained plain. To be specific, they have not grasped the advanced pronunciation techniques such as assimilation, liaison, stress, loss of plosion and incomplete plosion, let alone native-like pronunciation and intonation. The second is that they have not grasped communication rules. Results of the assessment from the traits of "Illoc" and "Socio" show that about two thirds of the candidates can express their ideas clearly and give timely and reasonable response to others' utterance. In addition, their utterance is appropriate to the actual language use situation. However, the results are not so satisfactory when it comes to flexible interaction. Nearly half of the candidates lack communication means. They cannot find appropriate time to join in a discussion. Some of them interrupt others whenever they like, some keep on talking when they get a turn, and some always wait for others to give them a turn. The third is that candidates generally lack nonverbal communication awareness. A great number of linguists and experts on language teaching and assessment keep arguing that nonverbal communication is indispensible to communicative proficiency (such as Harrison 1965, Morlan & Tuttle 1976, Canale 1983, Canale & Swain 1980, Ross 1986, Nunn 2000, Taylor 2006, etc.). It is a common feature of everyday communication among native speakers. However, about half of the candidates who underwent this assessment procedure do not show this ability. While presenting ideas, they display flat tones. While discussing with others, they do not make direct eye contact. Their facial expression is dumb. They are unable to use gestures. Even though some of them use some gestures, they are not natural. The above three problems are common among Chinese college students,