第二语言测试问题研究

柴省三 著

Research on Second Language Testing

对外经济贸易大学出版社
University of International Business and Economics Press

第二语言测试问题研究

柴省三 / 著

Research on Second Language Testing

对外经济贸易大学出版社
中国·北京

图书在版编目（CIP）数据

第二语言测试问题研究/柴省三著. —北京：对外经济贸易大学出版社，2018.8

ISBN 978-7-5663-1945-6

Ⅰ.①第… Ⅱ.①柴… Ⅲ.①汉语—水平考试—研究 Ⅳ.①H19

中国版本图书馆 CIP 数据核字（2018）第 150274 号

第二语言测试问题研究

Di-er Yuyan Ceshi Wenti Yanjiu

柴省三 著

责任编辑：史伟明

对外经济贸易大学出版社

北京市朝阳区惠新东街 10 号 邮政编码：100029

邮购电话：010－64492338 发行部电话：010－64492342

网址：http://www.uibep.com E-mail：uibep@126.com

北京九州迅驰传媒文化有限公司印装 新华书店经销

成品尺寸：170mm×230mm 17 印张 333 千字

2018 年 8 月北京第 1 版 2018 年 8 月第 1 次印刷

ISBN 978-7-5663-1945-6

定价：49.00 元

柴省三

男，教授，博士研究生导师，北京语言大学语言科学院学术委员会委员。主要研究方向为语言测试、第二语言习得研究、对外汉语教学研究的实验设计与统计分析。

主持和参与国家、省部级重大或重点科研项目及一般项目6项，校级重大基础研究项目和一般科研项目8项，参与和独立完成的著作3部。在《外语教学与研究》、《世界汉语教学》、《语言教学与研究》、《汉语学习》、《语言文字应用》、《西安外国语大学学报》、《云南师范大学学报》（对外汉语教学版）、《中国考试》、《中国远程教育》、《现代语文》、《中国教育信息化》以及《语文教学与研究》等期刊上发表学术论文近40篇。

电子邮件：cxs66@blcu.edu.cn

责任编辑： 史伟明
责任印制： 沈德军
封面设计： 春天·书装工作室

国家社会科学基金重大项目资助（项目批准号：17ZDA305）

前　言

语言测试的理论研究与开发实践，特别是第二语言测试研究与实践是应用语言学中发展最快的领域之一。基于现代教育测量理论的第二语言测试研究与实践虽然只有不到 60 年的历史，但其发展速度却是有目共睹的。20 世纪初，大规模、标准化第二语言测试研究始于美国；1964 年，世界上第一个针对英语作为第二语言学习者（ESL）的大规模、标准化英语水平考试（即 TOEFL）正式开考，标志着在英语作为第二语言测试的理论研究和应用实践两个方面，均已取得了令人瞩目的成就。

20 世纪 80 年代初，随着国内来华留学生人数的高速增长，我国针对汉语作为第二语言的水平测试研究也取得了重大进展。1984 年，北京语言大学（前北京语言学院）敏锐地意识到汉语作为第二语言的测试工具在语言水平测量、对外汉语教学标准构建和教材研发等方面的重要地位，正式启动了以测量母语非汉语者的汉语水平为主要目的的中国汉语水平考试（HSK）的调研和实验探索工作；1989 年，北京语言大学（前北京语言学院）正式组建了中国国内第一个以汉语作为第二语言测试的研发、实施和推广为主要目标的专业化考试机构（即汉语水平考试中心）。1993 年，作者有幸成为北京语言大学汉语水平考试中心的一员，从此对第二语言测试和习得研究产生了浓厚的兴趣。

近 15 年以来，作者一直对第二语言测试的效度与信度问题、语言测试的公平性（fairness）问题、常模参照性（NRT）语言测试和标准参照性（CRT）语言测试工具长度的决策理论与方法问题、语言测试的信息化问题、中美留学生教育考试体系的比较研究以及各种面向社会的语言测试研究等问题抱有浓厚的兴趣，并将针对上述问题的部分理论研究成果、实证研究成果陆续发表在《世界汉语教学》《语言教学与研究》《语言文字应用》《中国远程教育》《中国考试》《中国教育信息化》和《云南师范大学学报》等学术期刊上。受期刊版面的限制，很多研究成果存在系统性、完整性不足以及深度不够的问题，因此，作者决定以上

述研究内容作为基本线索，针对有关问题进行了更深入、全面的探讨，并从更多的角度进行了较大幅度的充实与完善；此外，为了保证专著的系统性，书中还有许多内容属于最新的、从未公开发表过的研究成果。

本书正文由七章构成，研究内容不仅注重对理论、方法的探索，更注重基于实测数据的实证研究。

第一章以北京语言大学汉语水平考试中心研发的中国汉语水平考试（HSK）的效度和信度问题为研究对象，针对阅读理解和听力理解测验的效度问题、考生的听力理解应试策略和 HSK 测验分数的历时稳定性等问题进行了实证研究。

第二章首先针对第二语言测试公平性的理论问题进行了探讨，然后以实测数据为基础，通过 DIF 检验方法和蕴含量表分析法（implicational scaling）分别对中国汉语水平考试（HSK）阅读理解测验的题目和篇章偏差及公平性问题进行了实证研究。

第三章首先以经典测量理论为基础对实用汉语水平认定考试（C. TEST）测验长度的合理性进行了分析，然后以概化理论为基础，对中国汉语水平考试（HSK）阅读理解测验长度的科学性问题进行了检验，最后，基于边际效应递减法则（LDMU）和概化理论对中国汉语水平考试（改进版）的测验长度问题进行了研究。本章的研究思路、方法和结果将有助于测验开发者、用户和其他读者正确地理解一个科学的语言测验究竟使用多少题目才算合理的实用性问题。

第四章专门对标准参照性语言测试的长度决策理论和方法进行了探讨。由于在标准参照性语言测试中所获得的测验结果通常与某个或某些外在的目标、领域或标准等直接关联，常模参照性测验（NRT）和标准参照性测验的潜在能力假设及分数分布要求并不相同。因此，本章在对标准参照测验的概念、特征、信度估计理论和方法进行简要考察的基础上，系统探讨了标准参照性语言测验长度决策的标准、过程与方法。

第五章研究的重点在语言测试的信息化问题。本章针对计算机自适应性语言测试的智能选题方法、中国汉语水平考试（HSK）考生信息资源库的建设与应用、汉语水平考试（HSK）远程 CAT 阅读理解测试模式以及 HSK 网络系统的使用现状等问题分别进行了理论探讨和实证研究。

第六章在对来华留学生的发展规模、学习层次和专业结构等特征进行详尽调研的基础上，专门对我国来华留学生教育招生考试体系与美国留学生教育招生考

试体系的基本框架与功能进行了对比研究，探讨了我国留学生教育考试体系中存在的缺失与错位问题，并对构建和完善我国的留学生教育招生考试体系提出了相应的对策、建议和具体的开发路径。

第七章针对语言测试的产业化发展趋势，对国内外面向社会的各种第二语言测试工具（也包括少数母语测试）进行了归纳、总结和分析。

书中的绝大多数研究内容都是围绕中国汉语水平考试（HSK）的信度、效度、公平性、考试的信息化和测验工具的长度等具体实践与理论问题而展开的。研究方法涉及经典测量理论（CTT）、项目反应理论（IRT）和概化理论（generalizability theory），研究目标力争紧密结合第二语言测试的实践问题和现实问题。作者希望研究成果能够为语言测试和语言习得研究领域，特别是将汉语作为第二语言的测试与习得专业的硕士研究生、博士研究生，对外汉语教师以及国内英语教学与测试研究者提供些许启发与借鉴。

特别值得说明的是，这本专著的出版过程实际上也是作者对自己职业经历和学术轨迹进行阶段性反思和梳理的过程。北京语言大学汉语水平考试中心是国内最早从事汉语测试的专业机构，具有良好的学术传统、研究氛围和独特的研究资源，但遗憾的是，由于本人禀赋不足，“熏”而无成。因此，书中的纰漏和不妥之处，恳请读者不吝赐教。

柴省三

2018 年 6 月于北京

目　录

第五章 语言测试信息化研究

第六章 中美留学生教育考试体系对比与研究

第七章 面向社会的语言测试研究

第一章

语言测试效度与信度研究

效度（validity）和信度（reliability）是语言测试领域中两个最重要的概念，也是评价语言测试分数解释和使用有效性以及测量误差大小的重要标准。如果语言测试缺乏效度，那么测验结果就不能充分反映被试者在拟测构想（constructs）上的实际水平；如果语言测试的分数缺乏信度，或者说信度不高，那么也就意味着测验工具的测量误差太大，考试分数也不能反映被试者的真实语言水平。效度和信度是教育测量和语言测试领域中永恒的话题，是测验的设计者、开发者、考试的用户和考生最关心的问题。本章首先以真实的测量数据为基础，对中国汉语水平考试（HSK）听力理解测验的构想效度问题进行研究；其次，在大样本抽样的基础上，从 22 582 名参加 HSK（初、中等）考试的考生中分别抽取低分组考生、中分组考生和高分组考生各 1 000 名作为研究样本，借助多元聚类分析法（cluster analysis）对汉语水平考试的阅读理解分测验（sub-section）的效度问题进行了研究；再次，基于问卷调查法对 HSK 的考生在听力理解测试中的应试策略（test-taking strategies）使用情况进行了实证研究，通过考察应试策略与考试结果之间的关系，为 HSK 听力理解测验的效度论证提供了必要的证据；最后，对中国汉语水平考试（HSK）的测验分数的复本（alternative-forms）稳定性进行了实证研究，为 HSK 考试的信度提供了比较有说服力的统计证据。

第一节　汉语水平考试（HSK）听力测验构想效度研究

听力理解能力是语言水平高低的重要标志之一，因此，作为测试母语非汉语者一般汉语水平为主要目的的中国汉语水平考试（HSK），主要从听力理解和阅读理解两个方面来测量考生的汉语语言能力。本节在对语言测试构想效度验证和测验方法进行探讨的基础上，使用等级聚类分析法（HCA）专门针对 HSK（初、中等）听力理解测验的项目维度（dimension）进行了研究。基于测验构想效度和任务真实性（authenticity）考虑，并参照听力理解测验项目的实证聚类结构与句子、简短对话、讲话或长对话三种测验任务的操作结构对应关系，对中国汉语水平考试（HSK）的听力理解测验任务形式的有效程度和改进方式，提出了自己的改进建议。

一、问题的提出

信度和效度是评价语言测验质量高低的核心指标，不过，信度具有一定的静态性，是效度的必要而非充分条件，信度的高低可以从测验分数本身获得刚性的佐证。效度（validity）则是柔性的，只能依据各种经验证据、理论证据或统计证据论证测验效度的高低。效度验证（validation）必须以测验分数的使用、解释或推断为参照点，依靠各种内、外部证据证明效度的高低。对于语言能力测验而言，最重要的是构想效度（construct validity）问题，即“语言测验实际测得的东西与理论假设的语言能力要素或心理特征相吻合的程度”（舒运祥，1999）。如果测验的构想效度不高，那么测验结果（测验分数或测验表现）就无法对被试者语言能力的高低做出有效的解释或推断（inferences），更无法准确地对考生在非测量情景中的语言表现水平做出有效的概括。

很久以来，关于效度，特别是构想效度的定义问题，在心理测量学界颇具争议。有不少学者主张将效度划分为内容效度（content validity）、效标效度（criterion validity）和构想效度等不同形式，但 Messick（1981，1989，1996）、Cronbach（1971）和 Bachman（1990）等人则认为，内容效度和效标效度是获得构想效度的手段，三者具有一元结构性。通过专家评判获得的内容效度，以及通过测验之间的相关分析所获得的效标效度，只是为构想效度提供证据和输入而已（Shohamy，1991），因此，在所有的语言能力测验中，构想效度问题都是测验设计、开发和使用中无法回避的首要问题。

在构想效度的验证过程中，获取证据的来源越广泛，证明测验使用或分数解释有效的说服力就越强。效度验证时所获得的各种类型的证据，包括重要的、但未必支持测验使用或解释的证据，都应该被测验开发者予以评估甚至采纳，以编制出更有效的测验或对现行测验进行必要的改进（modifications），使之更合理。证明测验使用或解释具有构想效度的过程是一个不断积累证据的动态过程。比如，我们可以通过考察测验实际所测的东西与某种理论依据（theoretical basis）是否吻合来说明测验是否有效，或者依靠专家的经验判断（judgmental reasons）从测验内容方面探讨效度的高低，也可以使用各种统计方法获得实证证据（empirical evidence）对效度进行验证。总之，效度验证不是一劳永逸的即时（synchronous）过程，而是一个使用多种方法从多方面、多渠道不断进行证据积累的历时性（diachronical）论证过程。

中国汉语水平考试（HSK）是以测试母语非汉语者的一般汉语能力为唯一目的的国家级标准化考试，其测量结果对考生求学、求职以及汉语能力评价的影响后效（backwash）均具有高风险性（high-stakes），因此，HSK 从设计之初就十分重视测验的效度问题。HSK 考试的基本理念是，以听力理解和阅读理解为主线，语言能力测试为核心，标准化形式（多项选择题）为基本方向（刘英林，1989）。为了比较准确地测量考生的一般语言能力，在测验任务设计时，要对考生的目标语言使用域及交际特征进行分析与概括，并从中选取若干典型样本作为测验任务。

考虑到听力理解能力测试在汉语水平考试中的特殊性，本书拟将使用实证统计法专门研究听力理解测验的构想效度问题。研究的基本思路是：首先运用多元等级聚类分析法对 HSK 听力理解测验的三种任务形式，即句子、简短对话和长对话（或讲话）三种听力测验项目所测量的潜在纬度（dimension）结构进行考察，然后验证测验项目纬度结构与 HSK 听力理解能力的一元结构构想是否吻合，最后以本研究的统计证据为基础，从提高测验任务的真实性（authenticity）角度出发，在不损害测验效度并维持现有分数体系的前提下，对 HSK（初、中等）听力理解测验任务类型的改进提出了自己的建议，以期研究结果为未来 HSK 测验方式的改进与完善提供借鉴。

二、听力测验的构想与测验方法

听（listening）是语言交际活动、第二语言习得的前提，是获取信息的重要途径之一，同时也是语言学习者生存、发展所必备的一项社会文化技能（Socio—cultural skill）（Anderson，1999）。在语言习得过程中，以听和读为主要形式的接受型（receptive）技能总是先于表达型技能（邹申，2005）。从语言能力结构角度来看，几乎所有的语言能力模型（models）都将听力理解能力看作模型的重要组成部分。听力理解能力被公认为是一个人语言水平高低的重要标志，在第二语言教学和测试中都占有重要的地位，因此，国内外几乎所有的综合性语言测验均以相当大的权重将听力理解能力列为必考的内容之一（如：TOEFL，IELTS，MELTB，TEF，J-TEST，CET，PETS 等）。不过，语言理解能力属于一种非常复杂的构想，因而对听力理解能力的认识是人类从心理学、语言学角度不断探索的漫长过程。从现有的 L1 和 L2 听力研究文献中，我们还没有找到一个测验可以依据的、被普遍接受的听力能力解释理论（explanatory theory）。在实践中，

测验设计者往往在参考其他相关测验的基础上，凭知觉或个人经验来编制听力理解测验（Buck，1991），因此，测验质量保证的理论依据是什么？或者说使用什么样的测验任务来测量听力理解能力问题，无疑是影响测验使用或分数解释效度的关键。

由于听力理解能力是一种抽象的特质（trait），它本身不能像物理测量那样进行直接测量，只能建立在对经验事实或行为观察的基础上（陈宏，1997）。因此，测验设计和开发者必须根据构想定义的能力要素，从那些与一般语言能力具有潜在对应关系的目标语言使用域（target language use domain）中，选择若干有代表性的听力材料作为测验任务（test tasks）输入，然后针对听力理解材料设计一个或若干个多项选择题，要求考生以对输入材料的理解加工为基础，从备选项目中选出正确答案。最后，再根据被试在测验任务上的表现情况，来推断其在目标语言使用域中的一般听力理解能力或特征。在应用性测量中，了解考生在特定测验任务上的表现，并不是测量所期望达到的最终目的，我们更感兴趣的是，把被试在测验上的表现或分数解释为个人语言能力水平指标（indicator）的有效性、合理性和恰当程度，而不是仅仅把被试的测验表现解释为其参加特定测验的应试能力（Bachman & Savignon，1986）。听力测试的有效程度就在于测试结果能否真正反映考生在现实生活（real-life）中应该具有的听力理解能力，因此听力测验任务的特征（characteristics）与目标语言应用任务特征的吻合程度，是影响测验构想效度的关键要素之一（Bachman & Palmer，1996）。如果两种任务特征脱节，那么考生在测验任务上的表现水平对考生在目标语言应用域中的语言使用表现就没有足够的概化能力（见图 1—1），即测验缺乏构想效度。

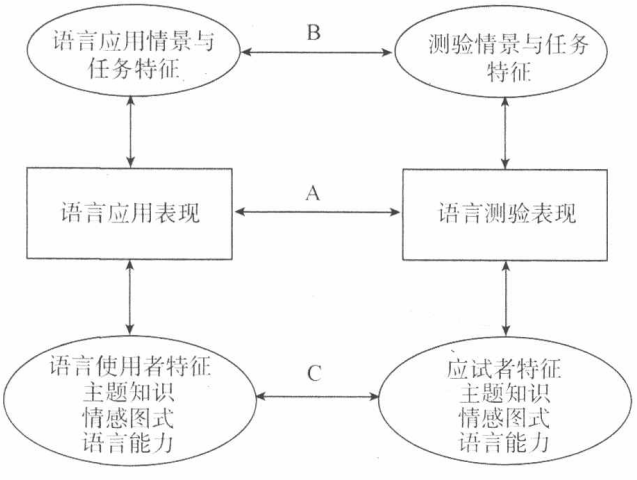

图 1—1　构想效度解释体系（引自 Bachman & Palmer，1996）

HSK（初、中等）的听力理解（listening comprehension）测验部分，在遵循 HSK 总体设计原则的前提下，主要承担对考生一般汉语听力理解能力的测量，即："通过测量考生能否听懂语速正常的句子、对话和一般题材的讲话（或长对话）来考察考生的听力理解能力"（刘英林等，1989）。测验所选用的语料在内容和题材等方面适当广泛。在语料集（sets）层面上追求多元性、平衡性和中立性（neutrality），避免语料在风格和内容等方面具有明显的专业倾向性，以便为不同地点、以不同方式和依据不同教材学习汉语的考生群体提供均等的考试刺激条件。

HSK（初、中等）的效度问题，一直是设计和开发者最关注的问题。不少研究人员（刘英林，1994）对 HSK（初、中等）设计的理念进行过最基本的探讨。张凯（1991）用因素分析法对 HSK（初、中等）测验的构想效度进行了研究，郭树军（1995）、陈宏（1997）分别用 Grant Henning 的内部结构效度法和多质多法（multitrait-multimethod）分别就 HSK（初、中等）的构想效度作过考察。这些研究所得出的结论比较一致，即：HSK（初、中等）考试，尽管其测验操作结构由 4 个分测验、9 个小部分（sub-sections）所构成，但从测验实际所测量到的能力因子来看，HSK（初、中等）主要测量了听力理解能力和阅读理解能力。HSK（初、中等）的听力理解测验构想就是考生的一般汉语听力理解能力。

三、听力测验项目的聚类分析

HSK（初、中等）的听力理解测验任务，是由基于三种不同类型的言语刺激形式所设计的 50 个多项选择题所构成，测验试题包括句子理解 15 题（用 S01 至 S15 表示题号）、简短对话理解 20 题（用 S16 至 S35 表示题号）以及长对话或讲话理解 15 题（用 S36 至 S50 表示题号）。测验的具体操作结构请参见图 1－2。通过三种相对独立的测验任务来测量考生能否了解句子、简短对话和讲话的基本大意，跳跃障碍，抓住主要信息或重要细节，根据所听到的材料进行推理和判断，以及对说话人的目的和态度的理解能力。本研究主要通过对测验项目进行聚类分析来探索句子理解测验任务是否测量了相对独立的因子。