目录

第一章 引言

语言是人类进行沟通和交流的表达方式，是人类思维的工具，也是文化的载体和传播工具。语言教学的主要任务是使学习者掌握语言交流所需要的知识和技能，培养跨文化交际的意识和能力，拓宽视野，形成正确的价值观。语言测试是语言教学的一个组成部分，目的是描述学习者语言能力的发展，评价学习者的语言能力水平，为语言教学提供反馈，助力语言学习，推动教学改革和发展。因此，狭义的语言测试反拨效应指语言考试或测试对语言教学和学习所产生的各方面影响。同时，语言的社会性是其本质属性。语言服务于社会，也随着社会的发展而发展。为了帮助机构或单位作出正确的录用、选拔、培训等决策，语言测试专业机构开发了各种类型的语言考试。这些考试具有特定的社会功能，用于描述和评价语言使用者的语言能力现状或发展潜力。语言测试的社会功能使其对考生个体、考试相关的利益群体乃至社会产生各种影响。因此，广义的语言测试反拨效应研究也涵盖了考试所产生的社会影响或后果。

本书对语言测试反拨效应的探讨主要围绕考试对教学产生的影响。然而，我国有不少大规模、高风险的语言考试，这些考试对考生、利益相关群体和国家的外语教育政策等都有着深刻的影响。因此，本书也将涉及考试产生的社会影响。本章是全书的引言部分，简要回顾语言测试反拨效应

研究概况，阐述相关的重要概念，并描述本书的编排结构，以帮助读者更好地阅读和理解本书的内容。

1.1 语言测试反拨效应研究概况

在我国历史上的科举制度实施期间，人们就认识到考试对教学产生的深刻影响。杨学为（2001：136）在探讨科举考试的历史评价时指出，“作为国家选拔官员的考试制度，科举的基本做法都是应当肯定的，‘肯定’当然不是说要恢复沿用，而是说，以历史唯物论的观点来看，它们并非错误。”杨学为从积极和消极两个方面分析了科举考试与教和学的关系。从积极方面来看，科举考试激励学子奋发学习，因为唯有努力才能考取功名；从消极方面来看，受科举考试内容和方式的影响，参加科举考试成为教学的目标，考试的要求成为教学要求，教学内容变得日益狭窄，教学成为考试的附庸，形成重理论、轻实践、死记硬背、呆板僵化的学风。

因此，考试制度会对教育产生影响并非新观点，但对此话题一直缺乏系统的研究。以高考为例，自1977年恢复高考以来，教学一直与高考密不可分。于涵（2017：128）在回顾高考制度恢复40年考试内容改革时指出，“除了选拔功能，高考因其连接高等教育与基础教育阶段主渠道的特殊地位，对于高中乃至整个基础教育的教学来说，都体现出了‘硬指挥棒’的作用，因此，高考也就自然被赋予了附加立场——引导教学。”如何发挥高考的正面作用，同时避免考试对教学产生的负面导向，依然是摆在我国教育工作者面前亟待解决的重大课题。

国外学者也早已关注语言测试对教学的影响。Wall（1997）详细介绍了语言测试反拨效应研究的发展历程。她引用Simon（1974）的报告，分享了教育测量领域长期以来对测试与教学关系的探索，研究内容包括考试

对学生的学习动力、期望、焦虑、学习方法的影响，对教学内容和方法的影响，对教学有效性评价的影响，对学校课程设置的影响，以及对社会机会分配的影响等各个方面。例如，牛津大学在1802年推出了一个新的考试项目之后，学生只关注该考试所考查的学科，这说明该考试项目限制了学生在校期间的学习内容。早期的语言测试领域也有零星的考试反拨效应研究论文或论著，如Wall（1997）介绍了三项早期的研究：Davies（1968）、Madson（1976）和Wesdorp（1982）。

在我们为撰写本书收集的语言测试文献中，最早聚焦教学与测试的是Ballard于1939年出版的专著《英语教学与测试》（*Teaching and Testing English*）。作者提出，测试是教学的重要组成部分，教学与测试之间很难划分出一条清晰的界限。作者以自己设计的一个朗读测试为例，说明测试对教学的影响。该测试要求学生在一分钟内尽可能多地朗读一连串单词，单词之间毫无关联且难度递增。这是一个考查学生词汇和语音知识的测试任务，但作者却多次发现教师在课堂上用该测试任务来训练学生，以提高学生的朗读速度。作者认为这是典型的考试误用，并用两个类比来解释自己的观点：父母通过不断给宝宝称体重来增加其营养，或是人们通过不断体检来增强自己的体质（Ballard 1939：161 - 162）。测试只是测量手段，不能替代课堂教学。

现代语言测试是应用语言学领域中的一个年轻学科。Lado于1961年出版《语言测试：外语测试的开发与应用（教师用书）》（*Language Testing: The Construction and Use of Foreign Language Tests: A Teacher's Book*）一书，标志着语言测试成为应用语言学的一个独立分支。在过去60多年的发展历程中，语言测试反拨效应受到越来越多的关注。20世纪90年代，英国兰卡斯特大学的测试学者基于在斯里兰卡开展的考试研究以及对相关研究的回顾，于1993年提出了著名的考试反拨效应是否存在之问（Alderson & Wall 1993；Wall & Alderson 1993），呼吁研究者应更加重视考试反拨效应的证据，更多地开展实证研究，并为今后的考试反

拨效应实证研究提出了15项假设。之后，语言测试领域的期刊《语言测试》（*Language Testing*）于1996年邀请Charles Alderson和Dianne Wall主编了语言测试反拨效应研究的专刊，发表了一批重要的研究成果，如Alderson & Hamp-Lyons（1996）分享了托福考试的案例，Messick（1996）提出了将考试后果作为效度组成部分的整体效度观，Bailey（1996）提出了考试反拨效应研究的理论模型等。之后，语言测试领域将考试反拨效应视为效度的一个重要组成部分，开展了深入的探索，涌现出一大批基于实证数据的调查和研究。本书第七章和第八章将挑选一些地区性考试和国际化考试的反拨效应研究案例进行分析。

1.2 语言测试反拨效应核心概念

每个研究领域都有一些核心概念，这些概念能展现学科的逻辑结构，对这些概念的阐释是对某个领域研究范畴的凝练和概括，可以帮助研究者抓住问题的要点，厘清事实依据，思考进一步探索的方向，同时也可以帮助读者正确理解和认识该领域的研究成果。在本节中，我们介绍以下几个语言测试反拨效应研究的核心概念：高风险考试，课堂测试，利益相关者，反拨效应、影响、后果。

1.2.1 高风险考试

语言测试根据其用途和产生的影响，可分为高风险考试（亦称高利害考试）和低风险考试。Madaus（1988：35）对高风险考试（high-stakes test）定义如下："此类考试的结果被用于重要决策，这些决策无论正确与否都将对学生、教师、教学管理者、家长或普通民众产生直接的影响。" Cizek（2005：25）认为，所谓高风险是指考试产生积极或消极的后果，如学生升级或留级、教师工资高低或奖惩、政府表彰表现突出的学校

或接管表现欠佳的学校等。Bachman & Damböck（2018：33）依据考试决策产生的风险程度将考试分为三大类：对考生产生重要影响的高风险考试（如高校入学考试）；影响程度中等的中风险考试（如语言课程中的分级测试）；影响程度较低的低风险考试（如课堂测试和反馈）。

Madaus（1988）对高风险考试的反拨效应提出了警告，他认为，考试的风险越高，对教学的影响越大。因为教师会为适应考试要求而改变教学目标和内容，最终考试要求很可能会成为真正意义上的课程标准。Cizek（2005）也承认人们对高风险考试有各种批评意见，如增加教师的挫败感和倦怠感、使低龄学生产生厌恶感、增加学生退学比例、减少课堂讲授时间、测试死记硬背的低层次内容、不利于追求学术卓越、缩减课程内容、拉大学业差距、助推社会不公平、引发作弊等。但他认为，这些批评往往来自考试相关者的主观评判或感受，并非基于令人信服的事实或充分的证据。而且，他从决策需求、问责制度等角度分析了高风险考试的作用，并且用数据说明了学生家长、普通民众等对高风险考试的认可。

我国是考试大国。隋朝建立的科举考试就是典型的高风险考试。科举制度之所以被废除，是因为它对教育、社会价值观、社会的改革和创新等产生了严重的消极影响。但是，科举考试制度依然“是中华民族对人类文明的伟大贡献，在世界上长期处于领先地位，许多国家借鉴经验，改革选拔制度，如英、法等国”（杨学为 2017：10）。新中国成立之后实施的高考是现代教育史上又一项重要的大规模考试，尽管高考也存在着片面追求高分的弊端以及由此产生的对教育的负面影响，但高考依然对公平、公正的人才选拔发挥着重要的作用。除了高考，我国大规模、高风险语言考试项目还有全国大学英语四、六级考试，全国高等学校英语专业四、八级考试和全国英语等级考试等。语言测试学者围绕这些考试项目，开展了多方位的、较深入的反拨效应研究。

1.2.2 课堂测试

“考试”与“测试”是语言测试领域最常用的两个词语，它们在很多语境中可以交替使用。尽管如此，两者在用法上仍有些区别。本书中，“考试”主要用来指风险程度相对较高的终结性语言能力测评；“测试”则用于风险程度相对较低的形成性评价。因此，教师在课堂上对学生的语言能力进行测评或学生自评和互评均被称为“课堂测试”（classroom-based assessment）。但是，当我们泛指对语言能力或水平的考查或评价时，“语言测试”是最为通俗的用法；而“语言考试”则用于特指大规模考试（见本书第七至八章）。从其发展历程来看，语言测试反拨效应研究主要聚焦大规模、高风险考试。Davies *et al.*（1999）编撰的《语言测试词典》（*Dictionary of Language Testing*）有600条术语，却没有包含“课堂测试”或“课堂语言测试”的相关条目。1984年创刊的《语言测试》早期收录的论文的主要研究对象是大规模考试或校本的终结性考试。课堂测试似乎是一个无需进行解释的普通名词，因为它是课堂教学的组成部分，几乎所有语言教师都在课堂上开展各种目的和形式的评价活动。

但是，随着语言测试领域对促学评价理念的深入探究，课堂测试以其独特的目的、方法和意义而得到越来越多的关注。2004年创刊的《语言测评季刊》（*Language Assessment Quarterly*）首期刊登了关于课堂形成性评价的论文（Leung 2004）。一些专业考试机构为更充分地体现从考试到测评的转向，修改了机构名称，如剑桥大学考试委员会（University of Cambridge Local Examinations Syndicate，简称UCLES）于2005年更名为剑桥大学考评院（Cambridge Assessment）。

课堂测试可以帮助教师获得大规模考试无法测量的学生语言能力信息，测试方式更加真实、有意义，是连接教、学、考最好的方式之一（金艳、孙杭 2020）。课堂测试主要用于课堂评估（classroom-based evaluation），是评价教学活动有效性、提升教学质量的重要手段。Genesee & Upshur（1996：5-6）对课堂评估的定义是，根据评估目的，

如确定学生掌握程度、了解学生学习困难、判断教学活动是否有效等，采用各种形式的课堂测试采集并解读数据，以便作出合适的教学决策。近年来，随着学习导向型评价（learning-oriented assessment）研究的深入，课堂测试也受到更多关注。学者在测试任务设计、教师角色、师生互动、促学效果等方面对语言教学中的课堂测试开展了深入的研究。

1.2.3 利益相关者

语言测试的利益相关者（stakeholders）指所有参与考试工作的人员、考试的使用者和受考试影响的个体或群体。Davies *et al.*（1999：185）对“利益”（stakes）的定义是“考试结果对考生产生影响的程度”。从考试反拨效应的角度看，受考试影响的不仅仅是考生，还包括教师、家长或亲属以及考试使用者，如高校或企业等选拔考生的单位，以及教育政策、语言政策甚至移民政策的制定者等。不同的群体对测试的需求不同，受影响的方面和程度不同，所需要的语言测评素养也不同。Taylor（2013）提出的语言测评素养剖面模型将利益相关者分为命题人员、课堂教师、行政管理人员和测试专业从业者四个群体。Butler *et al.*（2021）提出，语言测评素养研究还需要关注考生或语言学习者群体，他们是最重要的利益相关群体。

语言测试的社会学属性很大程度上源自考试所产生的利益。利益本身就是一个社会学名词，通常指得到的好处。语言测试的开发和使用是为了满足个体、群体乃至社会层面的各种需求。当测试相关者的利益诉求不同时，利益冲突便会产生。以获得公民资格或以移民为目的的语言测试为例（Shohamy 2009；Shohamy & McNamara 2009），考生希望通过考试实现身份转变或移民，国家权力机构则以语言考试为工具，实现对申请公民资格或移民的个体或群体的筛选，接受当权者认为对国家发展有利的考生，拒绝他们认为对国家发展无益或有潜在危害的考生。

1.2.4 反拨效应、影响、后果

语言测试反拨效应的研究范围很广，采用的术语也不尽相同。常用的几个意义相近的术语是“反拨效应/作用”（washback）、“影响”（impact）和“后果”（consequence）。

早期的反拨效应研究主要关注教学与考试的关系。在我国最早引荐标准化考试的《标准化考试——理论、原则与方法》一书中，桂诗春（1986：3）提出了一个简要的定义：一般来说，考试呼应教学，教学作用于考试；但是一些影响大的考试又往往反作用于教学，这种反作用被称为“反拨作用”。在该书的第一章第一节，桂教授分析了我国考试中存在的弊端，如考试缺乏明确目的，颠倒了教学和考试的关系，形形色色的应试辅导班、试题集、考试指导书泛滥成灾，导致学生疲于应付，干扰了教学的正常秩序；又如考试命题缺乏统一标准和要求，使考试无法对教学起到积极的“反拨”作用。桂诗春（1986：5）提出要建立符合我国人才培养需要的考试制度：所谓“制度化”，实际上指从教育和考试的相互关系上考虑考试的“反拨”作用，建立一套有利于促进教育发展的法定的考试制度。

考试对教学和学习产生的反拨作用是考试反拨效应研究最重要的内容。20世纪90年代初，Alderson & Wall（1993）提出了关于考试对教学和学习产生影响的15项假设，为系统化的考试反拨效应实证研究提供了思路和方法。随着语言测试使用范围的不断拓展，语言测试学者开始关注考试对教育体系或社会产生的影响，这种更具广泛意义的考试反拨作用被称为“影响”。Wall（1997：291）对语言测试的“反拨效应”和“影响”作了区分，前者被限定为考试对教学和学习产生的影响，后者则泛指考试在课堂、学校、教育体系乃至整个社会的范围内对相关的个体、政策或实践所产生的各方面影响。

“后果”一词主要源于Messick（1989，1996）提出的整体效度观。Messick（1989）认为，构念效度（construct validity）是考试效度最关键的层面，对考试分数的解释具有核心作用。影响考试效度的两大因素

是所测试的构念（construct）代表性不足或构念不相关。Messick（1989）采用渐进矩阵展示分层效度的框架，构念效度出现在每一个层面。他还在渐进矩阵中引入了价值含义（value implication）和社会后果（social consequence）。在此之前，考试效度研究更多关注考试本身，整体效度观则明确将考试产生的社会后果纳入效度论证的框架。在之后的效度论证框架中，社会后果成为效度研究的一个重要方面。Bachman & Palmer（2010）的测试使用论证框架（assessment use argument，简称AUA）更是将考试使用和后果作为效度研究的出发点。该框架提出四个有关考试效度的主张，其中的一个主张就是关于考试的使用及产生的后果，即语言测试的使用对所有利益相关者都是有利的。论证该主张的证据主要来自考试的教学反拨效应和社会影响。

因此，尽管这几个术语在语言测试研究中经常被交替使用，但是它们的意义和使用语境不尽相同。在本书中，我们用“反拨效应”来特指考试对教学和学习产生的作用，用“影响”或“后果”来泛指更广泛意义上的考试对教育体系或社会产生的影响。

1.3 本书结构

本书是“外语学科核心话题前沿研究文库 · 应用语言学核心话题系列丛书 · 语言测评”系列中的一册。全书共九章，从以下四个方面探索语言测试反拨效应研究的理论和实践：理论回顾与评析、研究方法介绍、实证研究案例分析以及发展趋势探讨。

第一至三章是语言测试反拨效应理论回顾与评析。第一章是全书的引言，简要回顾语言测试反拨效应研究概况，阐述相关的重要概念，帮助读者更好地阅读和理解本书内容。第二章回顾语言测试效度理论的发展过程，重点分析语言测试反拨效应与效度的关系。第三章回顾语言测试反拨

效应研究的理论框架，重点介绍几个具有较好应用价值的框架。第四至六章阐述语言测试反拨效应的研究方法。其中，第四章介绍定性研究方法，第五章介绍定量研究方法，第六章介绍混合研究方法。第七至八章首先介绍选择案例的主要原则和分析案例的框架，然后以具有一定影响力的地区性和国际化语言考试项目为例，回顾反拨效应实证研究的目的、方法和结果。第九章总结全书所阐述的主要观点和陈述的主要事实，分析语言测试反拨效应研究面临的挑战，探讨反拨效应研究发展趋势并提出选题建议。

需要指出的是，本书回顾和总结了语言测试反拨效应研究的理论和实践，但并未致力于拓展语言测试反拨效应的理论框架。例如，在第二章探讨考试反拨效应与效度理论的关系时，笔者回顾了相关的文献，提出所支持的观点，但是，本书并未深入探究效度理论框架，也没有提出自己的反拨作用研究理论框架。本书的主要目的有以下两个方面。第一，回顾国内外关于语言测试反拨效应研究的文献，指出加强反拨效应研究的重要意义；第二，界定语言测试反拨效应研究的范畴，梳理相关的研究范式和方法，推动我国的语言测试领域开展更多高质量的实证研究。反拨效应研究与教学息息相关，为更好地让语言类课程设计者、教材编写者和一线授课教师阅读本书，笔者尽可能采用通俗易懂的方式描述和论述相关的理论和方法。本书对语言测试反拨效应的理论和实践作了系统的梳理和回顾，因此也适合语言测试研究者，包括高校语言测试方向的硕博研究生，作为基础读物。