目　录

总　序……文秋芳 vii
前　言……顾永琦　王小英　张春青　李加义 x

第一部分　认识形成性评估

第一章　形成性评估的概念……2
1.1　形成性评估的定义……2
1.1.1　界定“形成性”……2
1.1.2　界定“评估”……4
1.1.3　界定“形成性评估”……7
1.2　形成性评估与相关概念辨析……9
1.2.1　常见相关概念……10
1.2.2　终结性评估与形成性评估……11
1.2.3　学习评估、促学评估及以评代学……12
1.2.4　另类评估……13
1.2.5　诊断性测评……13
1.2.6　过程性评估……14
1.2.7　自我参照评估……15
1.2.8　动态评估……15
1.2.9　面向学习的评估……20
1.3　形成性评估的工作模型……23
1.4　小结……27

第二章　形成性评估的理论依据……29
2.1　传统测试理论视角下的形成性评估……29
2.2　活动理论视角下的形成性评估……31

2.3 自主学习理论视角下的形成性评估 34
2.4 小结 36

第三章 外语形成性评估的内容 37
3.1 外语评估目标 37
3.1.1 语言能力构念 38
3.1.2 语言能力框架 43
3.1.3 课程标准 46
3.1.4 目标语言使用任务 47
3.2 外语评估标准 48
3.2.1 外语课堂任务的成功标准 48
3.2.2 构建语言标准 49
3.2.3 解释和使用标准 51
3.3 小结 52

第四章 形成性评估的方法 53
4.1 形成性评估收集证据的方法 53
4.2 形成性评估解读证据的方法 56
4.3 形成性评估提供反馈的方法 58
4.4 形成性评估跟进行动的方法 61
4.5 小结 63

第五章 形成性评估的质量标准和效度验证 64
5.1 语言测评的基本质量标准 64
5.1.1 Bachman & Palmer 的六个质量标准 64
5.1.2 Kunnan（2018）的测试公平性 68
5.1.3 形成性评估的质量标准 69
5.2 效度验证 73
5.2.1 传统的测试效度验证 73

5.2.2 传统的形成性评估效度验证.....74
5.2.3 基于论证的形成性评估效度验证.....76
5.3 小结.....84

第二部分 运用形成性评估

第六章 运用形成性评估的指导原则.....86
6.1 主体多样原则.....86
6.2 目标明确原则.....87
6.3 方法适宜原则.....89
6.4 基于标准和关注进步原则.....93
6.5 做出高质量反馈原则.....96
6.6 有后续教学活动原则.....98
6.7 小结.....99

第七章 课堂上的形成性评估.....100
7.1 词汇语法教学中的形成性评估.....100
7.1.1 学习过程不同阶段的形成性评估活动.....100
7.1.2 针对不同考查目标采用不同的评估工具.....101
7.2 听力教学中的形成性评估.....102
7.2.1 听力理解能力.....102
7.2.2 常用检测听力理解能力的工具.....102
7.2.3 听力课堂上不同阶段的形成性评估活动.....103
7.3 阅读教学中的形成性评估.....105
7.3.1 阅读理解能力.....105
7.3.2 常用检测阅读理解能力的工具.....106
7.3.3 阅读课堂上不同阶段的形成性评估活动.....106
7.4 口语教学中的形成性评估.....109
7.4.1 口语表达能力.....109
7.4.2 口语课堂上不同类型的形成性评估.....111

7.5 写作教学中的形成性评估 114
7.5.1 书面表达能力 114
7.5.2 写作课堂上不同阶段的形成性评估 115
7.6 其他方法 118
7.7 小结 120

第八章 课堂外的形成性评估 122
8.1 针对学生课外自主学习内容的形成性评估 122
8.2 针对学生课外自主学习的量的形成性评估 124
8.3 针对学生课外自主学习技能的形成性评估 127
8.4 小结 128

第九章 基于网络资源的形成性评估 129
9.1 自动评阅系统 129
9.1.1 自动评阅系统与形成性评估 129
9.1.2 iWrite 自动评阅系统 130
9.1.3 iWrite 自动评阅系统的使用 131
9.1.4 iWrite 提供给教师的反馈 134
9.1.5 iWrite 提供给学生的反馈 136
9.2 英语诊断性测评系统 138
9.2.1 英语诊断性测评与形成性评估 138
9.2.2 优诊学（高校版）诊断性测评系统 139
9.2.3 优诊学（高校版）给教师的反馈 140
9.2.4 优诊学（高校版）给学生的反馈 146
9.3 语料库 146
9.3.1 语料库辅助评判语言使用正确性 147
9.3.2 语料库辅助评判语言使用的恰当性 149
9.4 小结 150

第三部分 研究形成性评估

第十章 形成性评估研究概况.....154
10.1 形成性评估研究的关注点.....154
10.1.1 教师的形成性评估实践.....155
10.1.2 形成性评估的有效性.....157
10.1.3 其他研究.....159
10.2 形成性评估研究路线图.....160
10.2.1 从形成性评估概念框架看实证研究.....160
10.2.2 从效度验证框架到实证研究.....161
10.3 小结.....162

第十一章 形成性评估研究类型.....163
11.1 常见的研究类型.....163
11.2 研究问题.....166
11.2.1 提出研究问题.....166
11.2.2 基于课堂的形成性评估研究问题.....168
11.3 小结.....169

第十二章 形成性评估研究实例.....171
12.1 大学英语口语教师的形成性评估实践：一项个案研究.....171
12.1.1 文献综述.....172
12.1.2 研究方法.....174
12.1.3 研究结果.....175
12.1.4 讨论与结语.....184
12.2 形成性评估视角下的英语写作自动评阅反馈投入研究.....187
12.2.1 研究背景和研究问题.....188
12.2.2 研究方法.....190
12.2.3 研究结果.....192
12.2.4 讨论与结语.....200

12.3 研究实例简评 201
12.4 小结 202

第十三章 形成性评估研究展望 203
13.1 形成性评估的构念 203
13.2 教—学—评的目标 205
13.3 新技术背景下的形成性评估 205
13.4 形成性评估与大规模考试之间的关系 206
13.5 形成性评估的育人功能 207
13.6 小结 207

后记 209
参考文献 211
附录 239

总　序

“全国高等学校外语教师丛书”是外语教学与研究出版社高等英语教育出版分社精心策划、隆重推出的系列丛书，包含理论指导、科研方法、教学研究和课堂活动四个子系列。本套丛书既包括学界专家精心挑选的国外引进著作，又有特邀国内学者执笔完成的“命题作文”。作为开放的系列丛书，该丛书还将根据外语教学与科研的发展不断增加新的专题，以便教师研修与提高。

编者有幸参与了这套系列丛书的策划工作。在策划过程中，我们分析了高校英语教师面临的困难与挑战，考察了一线教师的需求，最终确立这套丛书选题的指导思想为：想外语教师所想，急外语教师所急，顺应广大教师的发展需求；确立这套丛书的写作特色为：突出科学性、可读性和操作性，做到举重若轻，条理清晰，例证丰富，深入浅出。

第一个子系列是“理论指导”。该系列力图为教师提供某学科或某领域的研究概貌，期盼读者能用较短的时间了解某领域的核心知识点与前沿研究课题。以《二语习得重点问题研究》一书为例，该书不求面面俱到，只求抓住二语习得研究领域中的热点、要点和富有争议的问题，动态展开叙述。每一章的写作以不同意见的争辩为出发点，对取向相左的理论、实证研究结果差异进行分析、梳理和评述，最后介绍或者展望国内外的最新发展趋势。全书阐述清晰，深入浅出，易读易懂。再比如《认知语言学与二语教学》一书，全书分为理论篇、教学篇与研究篇三个部分。理论篇阐述认知语言学视角下的语言观、教学观与学习观，以及与二语教学相关的认知语言学中的主要概念与理论；教学篇选用认知语言学领域比较成熟的理论，探讨应用到中国英语教学实践的可能性；研究篇包括国内外将认知语言学理论应用到教学实践中的研究综述、研究方法介绍以及对未来研究的展望。

第二个子系列是“科研方法”。该系列介绍了多种研究方法，通常是一本书介绍一种方法，例如问卷调查、个案研究、行动研究、有声思维、语料库研

究、微变化研究和启动研究等。也有的书涉及多种方法，综合描述量化研究或者质化研究，例如：《应用语言学中的质性研究与分析》《应用语言学中的量化研究与分析》和《第二语言研究中的数据收集方法》等。凡入选本系列丛书的著作人，无论是国外著者还是国内著者，均有高度的读者意识，乐于为一线教师开展教学科研服务，力求做到帮助读者“排忧解难”。例如，澳大利亚安妮·伯恩斯（Anne Burns）教授撰写的《英语教学中的行动研究方法》一书，从一线教师的视角，讨论行动研究的各个环节，每章均有“反思时刻”“行动时刻”等新颖形式设计。同时，全书运用了丰富例证来解释理论概念，便于读者理解、思考和消化所读内容。凡是应邀撰写研究方法系列的中国著作人均有博士学位，并对自己阐述的研究方法有着丰富的实践经验。他们有的运用书中的研究方法完成了硕士、博士论文，有的采用书中的研究方法从事过重大科研项目。以秦晓晴教授撰写的《外语教学问卷调查法》一书为例，该书著者将系统性与实用性有机结合，根据实施问卷调查法的流程，系统地介绍了问卷调查研究中问题的提出、问卷项目设计、问卷试测、问卷实施、问卷整理及数据准备、问卷评价以及问卷数据汇总及统计分析方法选择等环节。书中各个环节的描述都配有易于理解的研究实例。

第三个子系列是“教学研究”。该系列与前两个系列相比，有两点显著不同：第一，本系列侧重同步培养教师的教学能力与教学研究能力；第二，本系列所有著作的撰稿人主要为中国学者。有些著者虽然目前在海外工作和生活，但他们出国前曾在国内高校任教，也经常回国参与国内的教学与研究工作。本系列包括《英语听力教学与研究》《英语写作教学与研究》《英语阅读教学与研究》《英语口语教学与研究》《翻译教学与研究》等。以《英语听力教学与研究》一书为例，著者王艳副教授拥有十多年的听力教学经验，同时听力教学研究又是她博士论文的选题领域。《英语听力教学与研究》一书，浓缩了她多年来听力教学与听力教学研究的宝贵经验。全书分为两部分：教学篇与研究篇。教学篇中涉及了听力教学的各个重要环节以及学生在听力学习中可能碰到的困难与应对的办法，所选用的案例均来自著者课堂教学的真实活动。研究篇中既有著者的听力教学研究案例，也有著者从国内外文献中筛选出的符合中国国情的听力教学研究案例，综合在一起加以分析阐述。

第四个子系列是“课堂活动”。该系列汇集了各分册作者多年来的一线教学经验，旨在为教师提供具体、真实、具有较高借鉴价值的课堂活动案例，提高教师的课堂教学能力。该系列图书包括《英语阅读教学活动设计》《英语听力课堂活动设计》《英语合作学习活动》等。以《英语阅读教学活动设计》一书为例，阅读教学是学生学习语言知识和教师培养学生思维的重要途径和载体。该书第一作者陈则航教授多年来致力于英语阅读教学研究，希望通过该书与读者分享如何进行具体的阅读教学活动设计，探讨如何在课堂教学中落实阅读教学理念。该书包括三个部分。第一部分介绍在阅读前、阅读中和阅读后这三个不同阶段教师可以设计的阅读教学活动，并且介绍了阅读测评的目的、原则和方式。第二部分探讨了如何通过阅读教学促进学生思维发展。第三部分展示了教师在阅读课堂中的真实教学案例，并对其进行了分析与点评，以期为改进阅读教学活动设计提供启示。

教育大计，教师为本。“全国高等学校外语教师丛书”内容全面，出版及时，必将成为高校教师提升自我教学能力、研究能力与合作能力的良师益友。编者相信本套丛书的出版对高校外语教师个人专业能力的提高，对教师队伍整体素质的提高，必将起到积极的推动作用。

文秋芳
北京外国语大学
中国外语与教育研究中心

前　言

评估是教学必不可少的组成部分，教师时刻需要通过评估了解学生对教学目标的掌握程度，做出相应的教学决策，确保教学目标的实现。形成性评估是指教师、学生及其同伴获取、解读和使用学生学习结果的证据，并基于证据做出教学决策、改善教学实践的过程（Black & Wiliam 2009）。《普通高等学校本科专业类教学质量国家标准》中的《外国语言文学类教学质量国家标准》（教育部高等学校教学指导委员会 2018）和《大学英语教学指南（2020 版）》（教育部高等学校大学外语教学指导委员会 2020）都在评价部分强调了形成性评估的重要性，并提出具体要求。上述文件指出，评估应以促进学生学习为目的，应注重形成性评估与终结性评估相结合，选择科学的评估工具，合理解读和使用评估结果，及时提供反馈信息，不断调整和改进教学。

我国形成性评估研究方兴未艾。在中国知网以“形成性评估与英语教学”为主题进行搜索，限制条件为 2001—2021 年间在北大核心期刊上发表的文章，共搜索到 150 篇。文章涵盖不同学段、不同参与者的形成性评估与教学，涉及形成性评估的内容（口语、写作和阅读等）、手段（互评和档案袋等）、研究方法和辅助工具（现代信息技术）。从发表年份和数量来看，2005—2014 年构成发表数量的高峰期，发表数量最高值在 2008 年，为 16 篇；2014—2021 年间每年发表量在 4 篇以内。在教育部高等学校大学外语教学指导委员会（2018—2022）开展的全国范围的调研中（金艳 2020），高校大学英语课堂评估的主要方式包括教师课堂反馈（87%）、课后反馈（78%）、学生互评（51%）、自动评分系统反馈（41%）、学生自评（36%）以及其他方式（3%）。高校大学英语学习过程评估主要采用课堂表现评估（95%）、平时测验（90%）以及期中和期末考试（87%），7% 的高校报告未采用形成性评估。高校开展大学英语形成性评估的主要困难包括教师缺乏评估相关知识和能力（41%）、教师缺乏设计和实施评估的时间（54%）、学生不重视形成性评估（41%）、高校没有要求开展形成

性评估（19%）。以上调查结果表明，形成性评估已在高校大学英语教学中得到一定程度的重视，但依然存在教师形成性评估素养匮乏等问题。

从教学实践看，通过分析某次会议的 11 个国家一流课程发言和 9 篇教学改革文章及相关研究成果发现，“形成性评估”占总分数的比例在多所高校达到 50% 以上，有的甚至达到了 70%。但这些高校提到的“形成性评估”并不都是真正意义上的形成性评估。对形成性评估概念的误解主要体现在三方面。首先，许多教师难以分清形成性评估和终结性评估。例如，多所高校将小测、平台作业和课堂展示当作形成性评估。实质上，如果教师只是记录一个分数，而不是基于标准解读小测、作业和课堂展示所体现出的学生强项和问题，从而做出反馈并与学生一起改进教学和学习，那么这些小测、作业和课堂展示就只是一个个小型终结性评估，无法起到促学作用。反之，如果教师能利用学生期末考试的表现来改进下学期的教学，即形成性地使用终结性评估，也同样能够促进学生学习。此时，终结性评估也可以算作形成性评估的一环。其次，许多教师报告的形成性评估包含了与语言能力和教学目标无关的因素。最为常见的是考勤和线上学习时长，有的高校还将学生的努力程度纳入考虑范围。这类评估与语言能力无关，不能算是英语学习的形成性评估。由于这类评估发生在学习过程中，应该属于过程性评估范畴。将这些因素计入分数，虽有激励作用，但最终成绩很难体现学生的英语能力。最后，在众多报告中，教师只重视能产生记录的有形的评估，而没有提到课堂中时刻都在进行的即时形成性评估。有的高校在课堂表现部分提到对预习情况的评估、学生主动回答问题的记录以及依据评分标准对课堂展示的评分，但未提及教师在学生回答之后给予了怎样的反馈，更未提到教师和学生发现问题后该如何行动和改进。

无论是从调查数据，还是从教学实际看，教师在形成性评估的实施方面都存在评估素养不足的问题。评估素养是对教育评估的基本理解和应用相应知识对学生学习成果进行评估的技巧（Stiggins 1991）。评估素养越来越被认为是教师专业知识的有机组成部分，原因在于：(1) 评估，尤其是形成性评估，在学习中发挥核心作用，被证明能够显著促进学生学习；(2) 作为教育评估的执行者，教师在教育评估中发挥关键作用，很多教育评估改革措施的落地都需要教师具备较高的评估素养。为使学生获得较高学业成就，教师需要具备和掌握不

同种类和适应不同学习水平的评估素养。《大学英语教学指南（2020 版）》（教育部高等学校大学外语教学指导委员会 2020：31）专门指出：

> 为有效地开展评价与测试工作，大学英语教师需要提高自身的**评价素养**，教学管理部门需要加强对大学英语教师评价知识和技能的培训，特别是教学过程中的**形成性评价**理论和实践能力培训，使教师能够掌握**促学评价**的理念，采用**促学评价**的方法，处理好测试与教学的关系，更有效地利用评价与测试改进教学，切实保证大学英语课程教学的质量，实现大学英语课程的总体目标，满足国家和社会对大学生英语能力的需求。

一般来说，英语教师的评估素养包括知识（构念和测量等）、技能（命题和数据分析等）和原则（公平和促学反拨等）三个方面。例如，Fulcher（2012）把语言评估素养定义为：设计、命制、修改和评价大规模标准化考试和课堂测验时所需的知识、技能和能力；对测试过程的熟悉度；对指导和支撑实践的原则和概念（包括伦理问题和行为准则）的意识。Kremmel & Harding（2020）的研究发现了九个评估素养的构成因素，除了语言和语言发展外，英语教师比较关注的与评估有关的排前三位的因素是：(1）后效和备考，包括对课堂教学、教学材料、课程或大纲的设计的影响和备考等方面；(2）评估与教学，包括诊断学习者的长处和不足，以评促学、以评促教，反馈、自评和互评等；(3）评估原则和分数解释，包括测试的信度和效度、对分数与学习者学习能力的关系的解释等。

本书基于以上背景撰写，以期在满足教师实施形成性评估的需求，同时为教师研究形成性评估提供指引。本书分为对形成性评估的认识、运用和研究三部分：认识部分包括形成性评估的概念、理论依据、评估内容、评估方法以及质量标准和效度验证；运用部分包括形成性评估的指导原则、课堂上各语言技能的形成性评估、课堂外以及基于网络资源进行的形成性评估；研究部分关注形成性评估的研究概况、研究类型、研究实例以及未来展望。

虽然本书的目标读者是我国高校外语教师，但在分析形成性评估研究时，我们尽量囊括国内外相关研究，也没有完全局限于形成性评估在高校外语教学中的应用研究。这一方面是因为国内形成性评估研究数量有限，另一方面，跳出中国高校外语教学的环境看形成性评估会给读者一个更广阔的视野。因此，

本书从理论到实践到研究既适合广大高校外语教师，也能为硕博研究生，师范院校的高年级测评课程研发、管理和实施者，以及其他对形成性评估感兴趣的学者提供较为全面深入的参考。

在本书的筹备过程中，四位作者都参与了全书的整体策划和内容撰写，以下为具体撰写分工。前言（顾永琦、张春青、李加义）；第一部分：简介（李加义），第一章（王小英、顾永琦），第二章（王小英），第三章（李加义），第四章（王小英），第五章（顾永琦、张春青、李加义）；第二部分：简介（李加义），第六章（张春青），第七章（王小英），第八章（王小英），第九章（张春青）；第三部分：简介（李加义），第十章（顾永琦），第十一章（李加义），第十二章（王小英、张春青），第十三章（顾永琦）；后记（顾永琦）。顾永琦作为本书的统筹者，对全书进行了宏观把握和细节处理，为本书的连贯性和一致性提供了保证。王小英和张春青承担了全书大部分章节的撰写工作。此外，李加义还承担了全书的统稿和校订工作。

在此，我们特别感谢外研社高等英语教育出版分社副社长段长城。从选题立项到规划写作与统筹审校，她投入了大量心血，没有她的及时鼓励与鞭策，本书难以顺利出版。另外，审稿人的宝贵意见和建议使本书结构更加清晰、逻辑更加流畅、内容更加充实，让我们受益匪浅。本书责任编辑周娜细致认真的编校工作为本书提供了质量保证。在本书成书过程中，我们得到了很多同事、朋友和家人的帮助、支持与鼓励，在此一并致谢！

顾永琦
惠灵顿维多利亚大学
王小英
北京外国语大学
张春青
浙江外国语学院
李加义
澳门城市大学

2024 年 7 月

第一部分
认识形成性评估

本书第一部分介绍如何认识形成性评估，包括第一至五章。第一章介绍形成性评估的定义、形成性评估与相关概念辨析以及形成性评估的工作模型。第二章从传统测试理论、活动理论和自主学习理论分析形成性评估的理论依据。第三章梳理和探讨外语形成性评估的内容。第四章从形成性评估的实施步骤出发，介绍形成性评估的方法。第五章描述语言测评和形成性评估的质量标准，并阐述形成性评估的效度验证方法和效度论证模式。

第一章　形成性评估的概念

自 Black & Wiliam（1998a）发表形成性评估研究综述以来，形成性评估在教育体系中的重要作用引起日益广泛的关注。我国《义务教育英语课程标准（2022 年版）》《普通高中英语课程标准（2017 年版 2020 年修订）》以及《大学英语课程教学要求》等文件都对英语形成性评估提出了明确要求。近年来，国内外关于形成性评估的讨论和实证研究大量涌现，成为外语教育领域的研究热点之一。但是，究竟什么是形成性评估，目前学界尚未有统一公认的定义，有一些理解还存在一定的偏差（顾永琦、李加义 2020a；罗少茜等 2015）。从字面意思看，形成性评估是指为了形成性目的或具有形成性特点的评估。本章通过剖析形成性是什么、评估是什么、形成性评估具有怎样的突出特点，阐释形成性评估的本质，并澄清一些误解。

1.1　形成性评估的定义

1.1.1　界定“形成性”

20 世纪 60 年代，Cronbach（1963）提出，“评价是为决策提供信息的过程”，主张将评价放在教学或课程改革中，而非结束后，强调评价的“改进”功能。Scriven（1967）对此文做出回应，首次提出了“形成性”这一术语。他总结了评价的两大功能：用于评价项目最终结果的效果或价值（value）的终结性评价（summative evaluation），以及用于项目实施过程中做出改进项目的决定的形成性评价（formative evaluation）。他进一步指出，这两类评价可以基于相同的信息，并且应该同终结性评价一样仔细设计、认真实施形成性评价，而不是采用不正式的方式（Scriven 1991）。初期的形成性主要有两大特点：（1）目的，为了改进；（2）实施时间，在项目进行过程中而不是结束以后。同时期的 Stufflebeam（1966）提出了用于项目评价的 CIPP 模型，分为背景评价（context evaluation）、输入评价（input evaluation）、过程评价（process evaluation）和结果评价（product

evaluation）四部分，其中过程评价可被看作形成性评估的雏形。

Bloom *et al.*（1971：117）将形成性评价引入针对学生学习的教学领域，并将这一概念重新定义为“运用系统评价改进课程建设、教和学中的任何一个过程”。这里延续了形成性的目的和实施时间两大特点，只是将其从评价一门课程迁移到评价学生的学习。

20 世纪 80 年代以来，学者们开始关注形成性评估如何提升学生的学习效果（Andrade & Cizek 2010；Black & Wiliam 1998a，2003；Sadler 1989；Torrance 1993；Torrance & Pryor 1998），这个时期对形成性的界定也愈发强调其对提升学习效果的有益作用。这时的定义虽与传统定义一脉相承，但存在细微差别。传统的形成性只强调“为了改进的目的”，未体现是否要出现改进的效果。但这个时期对形成性评估的理解为：只有出现改进的效果，才是形成性评估。例如，Sadler（1989：120）认为，形成性评估关注的是“如何利用对学生作答质量的判断，通过缩短试错学习的随机性和低效性，塑造和提高学生的能力”。Black & Wiliam（1998b：140）指出，“当证据被用于调整教学工作以满足需要时，才是‘形成性评估’”。

21 世纪以来，学者们不再严格要求形成性评估务必带来改进的效果。例如，Black & Wiliam（2009：9）认为，“如果教师、学生及其同伴收集、解读并且使用关于学生学习结果的证据，并基于证据做出关于下一步教学的决策，且这样做出的教学决策有可能比不收集这些证据做出的决策好或者更有根据，则这样的课堂做法就是形成性的”。Black & Wiliam 指出，任何评估，只要是有可能带来有益效果的，就是形成性的。但是，形成性评估还须对教学产生一定影响，正如 Davison & Leung（2009：398）指出的，形成性评估“需要具有两大功能：提供信息（informing）和促成进步（forming）”，即形成性评估不仅需要为学生提供充分有用的信息，还需要将评估信息用于调整教学、促成进步。

Glaser（1963）提出的标准参照测量（criterion-referenced measurement）同样对形成性的概念产生了影响。它与常模参照测量（norm-referenced measurement）不同。Glaser 指出，在传统的常模参照测量中，测量结果用于评判学生在整个测试群体中的位置，这类测试的目的是排名和选拔，但不能很好地反映教学的有效性。因此他提出标准参照测量，目的是通过将学生实际获得

的知识和/或技能与预先设定的标准相比较，了解教学目标和标准的实现情况。

虽然 Glaser 最初将标准参照测量应用于课程评估中对学生成绩的解读，但是今天的旨在提升教学效果的形成性评估已经充分吸收这一概念。例如，两条被广泛认可的形成性评估策略是：“澄清、分享和理解学习目标及成功标准”和“向学生提供反馈帮助其进步”（Carless 2011：8；Wiliam 2010：31；Wiliam & Thompson 2008：64）。可以看出，形成性评估强调对学生表现的判断须以教学目标为参照，这样才能明晰学生应达到什么水平，目前处于什么水平，如何达到目标，而这正是形成性评估背后的教学理念（Black & Wiliam 2009）。正如 Popham（2011）所认为的，相较于常模参照测量，标准参照测量能更好地服务教师的形成性评估，因为依照标准对学生的表现做出的解读能为教师提供关于学生学习情况的更清晰的画面。

总体来看，形成性评估中的形成性具有以下特点：（1）在教学过程中实施；（2）关于学生学习而不是关于一个项目；（3）以既定标准为参照；（4）以改进教学为目的；（5）可能对随后的教学产生有益的影响。这些特点缺一不可。例如，教师在一个学期中定期进行小测验，检验学生对学习内容的掌握情况，这种做法虽满足形成性的时间特点，但评估目的只是了解“学生现在走到哪里了，而不是学生下一步往哪里走”（Torrance 1993：340），缺乏形成性的目的，因而这种测评不属于形成性评估，而是多次小的终结性评估（Popham 2011）。又如，一位教师认真批改学生作业并给出详细反馈，但随后并没有设计新的活动来检查学生是否吸收了反馈，而是开始一个新的单元教学。这位教师的做法虽然体现了提供信息的功能，但是其促成进步的功能缺失，即没有对随后的教学产生影响。换言之，只有同时具备上述五个特点，评估才有可能具有形成性。

1.1.2 界定“评估”

本质上，评估是为了某个评估目的而进行的推断过程，指基于一定的证据对人、事或属性进行的评判和估算，是包含如下基本步骤的实践过程：设计评估工具，使用评估工具收集相关信息，解读、评判信息，运用评判结果实现某种目的。传统测试也是这样的推断过程（McNamara 2000），因此，Brookhart

（2003）指出，评估是基于传统测试发展起来的。然而，由于其使用的情境与传统测试不同，评估逐步发展成为与测试相对应的一种新范式（Gipps 1994）。

自 20 世纪 80 年代以来，随着公众和政府对传统标准化考试的日益不满（Shepard 1989），作为评估者的一线教师（而非测试专家）的课堂评估做法受到越来越多的重视（Crooks 1988；Natriello 1987）。自 Black & Wiliam（1998a）的形成性评估研究综述发表以来，大量与形成性评估意思相近的概念不断涌现，如课堂评估（Brookhart 2004；Rea-Dickins 2001）、促学评估（Assessment Reform Group 1999；Stobart 2008）、基于教师的评估（Davison & Leung 2009）、形成性课堂评估（McMillan 2007a）、教师形成性评估（Leung & Mohan 2004）、替代性评估（Fox 2008）、基于课堂的语言评估（Rea-Dickins 2007，2008）等。这使得评估一词日益流行，也使评估与传统测试之间的联系与区别日益鲜明。

评估是包含与测试类似的基本步骤的实践过程。例如，Cizek（1996）指出，课堂评估是一个有计划的过程，包含收集并整合相关信息以发现学生的优点与不足、计划并改进教学、评价学生进步，并做出相关决定。McMillan（2007b）在其教材中给出了类似的定义，课堂评估是指收集信息、评价信息并运用信息帮助教师做决定的过程。Black & Wiliam（2009）依据对形成性评估长期的研究和实践，给出了更为全面的定义（详见本章第 1.1.1 小节）。这些定义反映出评估包含的基本步骤：收集证据（收集关于学生学习情况的信息）、解读证据（依据一定的标准对收集到的信息进行整合、评价和解读）、使用解读后的信息（如反馈学生的优点或不足、调整随后的教学计划等），与测试的步骤基本一致。

虽然评估与测试的步骤类似，但由于评估一般用于课堂教学领域，而非课堂之外的大规模标准化考试，因此在目的、内容、参与者以及具体步骤上，评估都体现出有别于测试的特点。测试的主要目的是为了做出决策，如中考、高考、雅思、托福，以及全国大学英语四、六级考试等，考查考生某些方面的能力，测试结果影响考生能否进入心仪的学校或者找到理想的工作。评估的主要目的是为了服务教学，如教师通过课堂提问或组织课堂讨论了解学生的学习情况及其与教学目标的差距，对学生表现做出的评判会影响随后教师的教学和学生的学习情况。通常，大规模、高风险的标准化测试由专门的测试专家组进行设计，在规定的时间、地点由专门机构负责实施，以确保测试的公平性，由经

验丰富的评分专家进行评阅，管理者和政策制定者将成绩排序并决定通过或录取分数线。课堂教学领域的评估通常由教师本人设计，有时学生也参与其中，其实施的时间、地点相对灵活，教师既可以提前设计，在计划的时间实施，也可以在上课过程中依据自己的观察判断随时实施（Cowie & Bell 1999）；时间上既可能在几秒钟之内完成，也可能需要一个学期才能完成整个过程（Wiliam & Thompson 2008）；既可以针对全班，也可以针对个别学生；教师主要依据课程目标对学生的表现进行评估；评估结果主要用于改进教师的教学和学生的学习。因此，Gipps（1994）提出，区别于传统的常模参照的、标准化、终结性测试，评估是一种基于标准的、形成性的、旨在改进教与学的新测评范式。虽然之后少有学者就这两种范式进行讨论，但越来越多的语言测试书籍、课程或会议都采用语言测试与评估（language testing and assessment）这样的说法，表明越来越多的学者认可这是两种不同的范式，是应用于不同情境、为了不同目标、地位同等重要的两种范式。Gu（2021a）对这两种范式进行了全面的对比（见表 1.1）。

表 1.1　语言测试与语言评估的差异（Gu 2021a：10-11）

	语言测试	语言评估
范式	心理测量的	基于标准的
目的	以结果为导向	以过程为导向
	终结性	形成性
	学生在未来能做什么？	学生到目前为止学到了什么？
	将学生排序	回答问题，如：检查学生的学习进展，诊断学生的问题
	考查语言水平	考查学习成就
	选拔、授予证书或做出决策	促进学生学习
	高影响	低影响
内容	基于需求分析或理论构念	基于课程大纲和教学目标
	更笼统	更详细、具体

（待续）

（续表）

	语言测试	语言评估
范围	规模较大	规模较小：班级、年级或学校层面
形式	更加标准化，形式固定	更加个性化，形式自由、开放
评分/分析	对结果进行统计分析	对结果进行质性分析
设计者	测试专家	教师
使用者	管理者，政策制定者	教师和学生
解读	常模参照	标准参照
	更客观	更主观

评估和测试本质上都是进行推断的过程（Bennett 2011），因此，传统测试理论的一些基本原则（如信度 [reliability]、效度 [validity] 等）也同样适用于评估（Brookhart 2004）。但由于评估使用的情境与传统测试有显著不同，评估逐步发展了自己的属性，成为区别于传统测试的新范式。

1.1.3　界定“形成性评估”

综合对形成性和评估的界定，形成性评估指包含一定基本步骤的具有形成性特点的实践过程。Bennett（2011）在其关于形成性评估的文献综述中指出，针对形成性评估这个概念存在工具论与过程论的分歧。持工具论者认为学期中实施的具有诊断功能的各种测评即为形成性评估。目前，我国中学教育中每门课都配备大量的单元或课程练习题、练习册，这正是工具论的典型反映。罗少茜等（2015：24）将这类评估定义为“内置于课程中的形成性评价”，与“即时形成性评价”和“计划互动性形成性评价”并列，同时指出这类形成性评估“实际上指的是为形成性目的开放的信息收集工具，类似于 ETS 开发的形成性评价题库”（罗少茜等 2015：25）。持过程论者认为，形成性评估是一个在教学中由教师和学生运用的实践过程，这一过程能对正在进行的教学给出反馈，提

升学生的学习成绩。Bennett（2011）认为两种观点都存在偏颇，只有将两者结合起来，才能使形成性评估真正实现形成性的功能。笔者赞同过程论的认识，因为形成性评估本质上是评估，是为了形成性的目的进行的包含几个基本步骤的推断过程（参见本章第 1.1.2 小节的讨论），上述工具论中提及的具有诊断功能的中期测验其实是收集证据这一步骤中可能用到的工具。将测评工具等同于形成性评估是对概念的一种误解，例如，我们不能将一份单元测试题当作评估，只有当教师通过测试题了解了学生的答题情况，并据此做出评判、给出反馈并改进随后的教学，完成一个完整的流程，才算评估活动。换句话说，评估是发生在具体的时间、地点，由特定人群参与的一项实践活动，一份试题只是评估活动中收集学生学习证据的一个潜在工具。

目前，常见的形成性评估定义（如 Black & Wiliam 2009；罗少茜等 2015）一般从评估目的、参与者、基本实践步骤以及与课堂教学的关系这些方面进行概念界定。Black & Wiliam（2009）的定义（详见本章第 1.1.1 小节）回答了形成性评估目的（改进教学）、参与者（教师、学生及其同伴）和基本步骤（收集、解释、使用信息）这些基本问题，并指出形成性评估其实是一种课堂实践（classroom practice），点明了形成性评估与课堂教学的紧密联系。罗少茜等（2015：13）将形成性评估定义为：

> 形成性评价是一种以评价为导向的课堂活动范式，它以评价者的判断能力为核心，要求评价者（教师、学生）采用、调整、设计各种适当的任务（课堂提问、任务、纸笔测试、档案袋等），系统地收集学生信息（包括学习产品和学习过程），并用适当的评价工具（检查表、评分准则等）对信息进行评价分析和阐释，再反馈给评价者（教师、学生）用于调整教和学的过程，促进学生语言能力的发展。

这个定义同样涵盖了评估目的、参与者和基本步骤，并且在基本步骤方面，将“使用”这一步骤细分为“反馈”和“调整教和学的过程”两步，更好地实现形成性评估需要具有的提供信息和促成进步两大功能（Davison & Leung 2009）。此外，这个定义提出了“形成性评价是一种以评价为导向的课堂活动

范式”这一观点，这是因为“形成性评价借用、引入、强化了许多评价测试中特有的概念和常用工具……使教学实践朝着更精确化、更科学化的方向发展”，因此是“连接评价与教和学之间的桥梁，使二者在课堂层面融为一体”（罗少茜等 2015：13）。虽然好的形成性评估与好的教学密不可分，但是好的教学并不等同于好的形成性评估（Bennett 2011）。为了更好地把握这一概念，形成性评估的定义应强调评估在本质上是一个推断过程，而不是落脚于课堂活动。

罗少茜等（2015）的定义提出形成性评估“以评价者的判断能力为核心”的观点，旨在强调评价者本身的评估素养对形成性评估的质量起着决定性作用。该观点认为，好的形成性评估需要教师掌握三方面知识：学科知识、教学知识和测评知识（Bennett 2011）。但这一观点是从教师的评估素养角度出发的，不是关于形成性评估本身。其实，许多研究者已经指出，形成性评估应以评估目标为核心，并在评估的每一个阶段都以评估目标为参照（如 Cowie & Bell 1999；Davison & Leung 2009；Wiliam & Thompson 2008），这是因为形成性评估本质上是探究学习者当前水平与目标水平之间差距的问题（Black & Wiliam 2009），明确的评估目标是形成性评估的出发点和最终归宿，并且对评估过程中的每个环节起到重要的指导作用。因此，整个评估过程都应围绕评估目标展开。只有评估目标与教学目标一致时，形成性评估才能服务于教学。

综上所述，我们将形成性评估定义为：在教学过程中实施的、为了改进教与学的推断过程，整个过程围绕评估目标展开，包含四个基本步骤，即收集学生学习证据、解读证据、提供反馈和实施跟进行动。

1.2　形成性评估与相关概念辨析

形成性评估的概念经历了几十年演化和发展，其间，学界在形成性和评估等不同的维度上形成了不同的理解，也在不同领域或范围内演化出了很多相关概念。另外，在向我国介绍的过程中，一些概念的翻译也比较杂乱。本节将对形成性评估及其相关概念做一个简单的梳理与辨析，对主要术语的翻译提出建议（见表 1.2）。

表 1.2　术语翻译建议

英文术语	中文翻译
assessment	测评（广义）；评估（狭义）
testing	测试
examination	考试
evaluation	评价
measurement	测量；量具
summative assessment	终结性评估
formative assessment	形成性评估
assessment of learning	学习评估；对学习的评估
assessment for learning	促学评估；促进学习的评估
assessment as learning	以评代学；作为学习的评估
alternative assessment	另类评估；替代评估
diagnostic assessment	诊断性测评
interim assessment	过程性评估；中期评估
ipsative assessment (self-referential feedback)	自我参照评估
dynamic assessment	动态评估
learning-oriented assessment	面向学习的评估；以学习为导向的评估

1.2.1　常见相关概念

测评或评估（assessment）有两层含义，可在广义上泛指一切测评，是个涵盖术语；也可在狭义上特指与测试所对应的评估，尤其指教学过程中教师与学生为改善教学与学习对学习效果所进行的评估。评估通常在教学过程中随时进行，因此一般不要求精确评分，但对教学与信息的判断也应该尽量准确。

测试（testing）一般指测试题的设计、使用和分析过程，也指研究该过程的学科。语言测试是指收集被试语言知识能力等信息，从而推断其语言知识的掌握情况及语言运用能力的过程或学科。

考试（examination）指考试题或考试过程。

评价（evaluation）指对某物的数量、质量或价值做出判断与评估，包含对价值的判断。

测量（measurement）既是动词测量（to measure）的名词形式，也可以指用作测量的工具（量具）。测量是对某物的精确度量，因此测量的工具、方法与分析要做到科学、严谨和精确。日常生活中的长度、重量、温度等需要精确测量，同样，对教育结果与心理概念的精确把握与研究也需要测量。

1.2.2　终结性评估与形成性评估

确切地说，终结性评估（summative assessment）与形成性评估（formative assessment）不是两种不同的评估形式，而是教育评估的两种不同目的。形成性评估发生在一个阶段之后，如期末，用以总结特定时间内学生所学到的技能和知识。评估者运用终结性评估的主要目的和关注点通常集中在评估的结果，比如学习成绩存档或评比。终结性评估通常用于成绩排名、资格认证、向学生家长报告或对教育提供者进行问责。形成性评估则随时发生在教学过程之中，教师与学生运用形成性评估时所关心的是学生学会了什么、没学好什么以及如何才能进一步提高。只有及时提供和获得有关信息才能及时利用这些信息改进教学与学习。

同一种形式的评估可以为不同目的服务。以同样方式收集的相同信息，如果用于帮助改进教学和学习，则其目的是形成性的，而如果只用于登记、存档或汇报，则其目的是终结性的。客人评价饭菜的咸淡是终结性评估，而厨师自己尝尝咸淡从而调整咸度则是形成性评估（Dirksen 2011）。

1.2.3 学习评估、促学评估及以评代学

形成性评估与终结性评估的说法容易被人误解为评估的性质或种类而不是评估的目的，于是便有学者寻求其他字眼取而代之。1999 年，英国的评估改革小组（Assessment Reform Group 1999）开始使用学习评估（assessment of learning，简称 A*o*L）取代终结性评估，并用促学评估（assessment for learning，简称 A*f*L）取代形成性评估。学习评估的目的是验证学生的学习情况，而促学评估的目的是促进学习的进步。学习评估通常发生在单元、期中或学期结束时的特定关键点，可用于对学生进行评分或排名，因此几乎等同于终结性评估。促学评估通常发生在整个教学过程中，会收集学生的理解和学习证据，并利用此类证据改进教学与学习，因此也称为形成性评估。由于学习评估也是对教学的评估，促学评估也能改善教学，因此本书对终结性评估与学习评估不做区分，也把形成性评估等同于促学评估。

以评代学或作为学习的评估（assessment as learning，简称 A*a*L）是个相对新鲜的概念。虽然该概念的雏形可能以前就有了，但真正提出该概念并将其与学习评估和促学评估并列而谈的则是 Earl（2003）。评估即学习的概念强调学生在学习过程中自我监控的重要性，使学习者对自己的学习负责，运用自我评估与反馈来探索如何改进后续的学习。作为学习的评估可以看作形成性评估的一种，它把重点放在学生身上，在过程上强调反思和自我评估以及学习者之间的协作，在结果上强调自主学习能力的培养。二十年来，虽然作为学习的评估的概念广受推崇，然而也很容易出现对该概念的错误理解，而且一旦在课堂中实施不当，会造成难以估量的损失。既然评估就是学习，那么课堂上可能会出现教师很少讲解，给学生留下的学习时间很少，把大量时间留给评估，认为评估了就是学习了的情况。这种情况长此以往会对学习产生非常不利的影响。因此我们同意 Wiliam（2018：683）所说的“评估不是学习”的观点，认为将二者区分开更加有利于分析评估的促学功能。