形成性评估的效度

顾永琦 李加义
新西兰惠灵顿维多利亚大学

外语教学

提 要：测评是教育体系中的重要环节，有效的形成性评估对于促进教学并最终影响素质教育的实施具有深远意义。但由于在形成性评估概念理解和应用上的偏差，难以达到理想的促学效果。本文首先强调明确形成性评估概念对形成性评估效度的重要性，其次详细阐述如何运用传统的语言测评质量框架和效度概念衡量形成性评估的质量，以期为教师的形成性评估实践和学生的学习能力提升提供指导和借鉴。

关键词：形成性评估；效度；外语教育
[中图分类号] H09 [文献标识码] A [文章编号] 2096-6105(2020)03-0034-08

1. 引言

教育测评历来是教育体系中不可或缺的重要一环，不仅能帮助教师获取教学信息、改进教学方法、保证教学质量，还有助于学生优化学习策略、改善学习方法、提高学习能力。教育测评既包括以标准化考试为代表的终结性测评（summative assessment），也包括以学习为目的、注重学习过程的形成性评估（formative assessment）（Leung & Mohan 2004）。

学生的具体学习情况信息主要来自课堂评估，并非标准化测试（Brookhart 2003；Earl 2003）。然而，多年来，作为教育改革和进步的重要推动力，具有较强甄别和选拔功能的终结性测评占据教育测评的主导地位（Cizek 2009；Shepard 2013）。20世纪60年代，Cronbach（1963）提出，测评是为决策提供信息的过程，主张将测评放在教学或课程改革中，而非结束后，强调测评的改进功能。Stufflebeam（1983）提出的CIPP模型由背景测评（context evaluation）、输入测评（input evaluation）、过程测评（process evaluation）和结果测评（product evaluation）四部分组成，其中过程测评可被看作形成性评估的雏形。“形成性”这一术语最早由Scriven（1967）提出，后被Bloom（1968）引入教学测评领域。随后，Bloom *et al.*（1971）重新定义了形成性评估和终结性测评，为课堂形成性评估的发展奠定了基础。但之后的20多年中，形成性评估并未得到足够的重视。直到Black & Wiliam（1998）发表关于形成性评估的研究综述，才在学界引起巨大反响。其主要结论是形成性评估对学生的学习具有深远影响。由此关于形成性评估的研究和实践在世界范围内迅速普及。

随后，Black & Wiliam（2009）综合已有研究，并依据他们对形成性评估的长期研究和实践，将课堂形成性评估定义为教师、学生或同伴收集、解释和使用学生学习情况信息的过程，以便教师作出改善教学的决策。形成性评估与教学紧密相关，其促学作用受到越来越多的重视，因此亦被称为“学习性评估”（assessment for learning）（Berry 2008；Berry & Adamson 2011；Broadfoot & Black 2004；Stiggins 2002，2005）。Gu（2020）将形成性评估的概念具体化为两种形式：理想的形成性评估和基本的形成性评估。前者由明确评估目的、完成评估实践和实现评估效果三部分组成（见图1），后者则需包含完成至少一个循环的形成性评估实践。在现实课堂教学中，理想的形成

性评估很难找到。形成性评估实践是由收集证据（elicitation）、解释证据（interpretation）、提供反馈（feedback）和后续行动（action）四个步骤组成的循环，只有构成完整的循环，才有可能使学生更靠近学习目标。形成性评估往往要经过多次完整的螺旋式循环才能帮助学生实现学习目的。

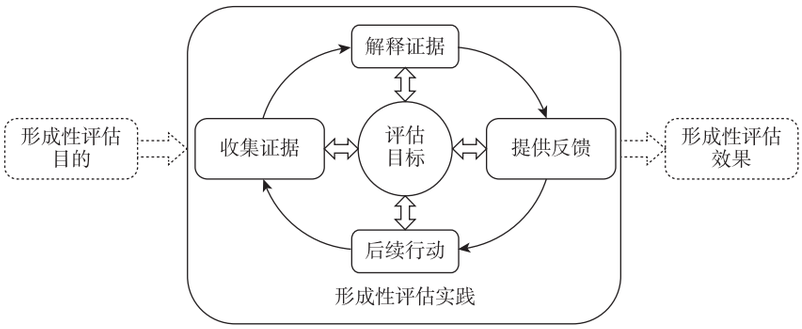

图1　形成性评估的概念具体化（Gu 2020）

在国内，形成性评估这一术语最先从教育学视角提出（卢耀增 1987）。在外语教育领域，中野照海等（1989）最早引介了Scriven（1967）提出的形成性评估概念。21世纪初，《义务教育英语课程标准》《普通高中英语课程标准》和《大学英语课程教学要求》等国家教育文件对英语形成性评估提出要求。自此，形成性评估研究开始引起国内外语教育研究者和实践者的广泛关注，逐步成为国内外语教育领域的研究热点之一。

近年来，国内外语教育领域中的形成性评估研究发展迅速，研究类型和研究主题愈发丰富。研究涉及的技能教学多样，涵盖听力（如夏晓燕等 2019）、阅读（如文秋芳 2011）、口语（如田朝霞 2018）、写作（如曾永红、梁玥 2017）、词汇（如毕鹏晖 2017）、翻译（如曹荣平、陈亚平 2013）等。相关研究涉及的教育层次包括中小学（如顾永琦、余国兴 2018）、高职（如李雪莲 2016）、大学本科（如张荔 2017）和硕士（如夏晓燕等 2019）阶段。研究方法包括综述性探讨（如黄剑等 2019；袁树厚、束定芳 2017）和实证研究（如李传益 2015；杨华、文秋芳 2014）。研究主题以形成性评估的实践方法和系统（如文秋芳 2016）为主，评估素养研究（如唐雄英 2017）和评估效度研究（如李清华、孔文 2015）较少。就评估方式和手段而言，几乎所有研究都涉及学生自评、同伴互评、教师评价等评估方式（如范劲松、季佩英 2017），大多数研究包括档案袋促学的实践手段（如罗少茜、张帅 2019；王慧文、刘芹 2018）。随着《中国英语能力等级量表》的问世，量表与形成性评估相结合成为形成性评估研究领域的新课题（如刘建达 2019；潘鸣威、吴雪峰 2019）。

近几年国内出现的外语教学形成性评估研究论文体现了我国外语教育领域形成性评估研究的最新现状，从中可以看出相关研究取得的进展和发展前景，但纵观整个研究领域，在形成性评估的概念解释和应用方面还存在诸多问题。第一，当前的文献研究对形成性评估的概念存在不同理解，造成一定偏差。首先，形成性评估概念理解方面的最大问题是假定所有形成性评估都能促进学习。实际上，考虑到形成性评估的质量，并不是所有的形成性评估都能改善学习。换句话说，良好的形成性评估可能改善学习，而不好的形成性评估则不能。其次，将形成性评估的操作方法（如教师评价、学生自评、同伴互评、多次评估等）和手段（如档案袋等）等同于形成性评估。最后，将形成性评估的目的或某个组成部分等同于形成性评估。例如，有些研究认为只要是以促进学习为目的的评价即为形成性评估，也有研究认为多测几次（多次收集证据）即为形成性评估，还有研究认为教师向学生提供的反馈即为形成性评估。第二，评估目标不够明确和具体，成功的标准模糊。若据此判断学生的成绩，其可行性则较低。第三，形成性评估实践循环不完整，步骤缺失。

将形成性评估中的某个组成部分等同于形成性评估可能直接导致形成性评估实践循环的不完整。有些研究强调目标的重要性，在评估实践中完成了收集证据、解释证据和提供反馈三个步骤，但鲜有研究认识到教师和学生应根据评估结果采取后续行动。只有反馈信息被用于改进学习时，反馈才能发挥形成性作用。

由于存在概念理解和应用上的偏差，形成性评估难以达到理想的促学效果。因此，明确形成性评估的概念是根本，也对形成性评估的效度至关重要。长久以来，效度(validity)是衡量大规模考试质量的核心标准，那么，形成性评估的质量标准是什么？如何确定？效度概念是否依然适用于形成性评估？本文将介绍形成性评估的目标、标准和效度，为教师和研究者的形成性评估实践提供启示。

2. 评估效度

效度理论经历了长期的发展变化，但争论至今仍没有结论(Newton & Baird 2016)。本文并不对效度理论进行任何补充，而是简要介绍效度概念的基本内核，为一线教师提供思路，使其对形成性评估质量标准的确立有清楚的认识。

2.1 教育测评的质量标准

效度是所有测评的中心议题(American Educational Research Association *et al*. 2014)。提及教育测评的质量，效度是重心中的重心、标准中的标准。当然，信度(reliability)也是测评质量不可或缺的重要标准。

传统上，效度指一项测试是否测量了它想测量的内容(Lado 1961)。这就是学界常说的内容效度(content validity)，即测试包含的内容是该测试应该涵盖的内容，因此也被称为“内部效度”。检验效度的最原始方法是计算该测试结果与已知标准的相关系数，也被称作“外部效度”。另外，效度也体现了一项测试与其理论构念相符合的程度，即构念效度(construct validity)。Cronbach (1988)提出理解效度概念的两个视角：测量视角(measurement perspective)和功能视角(functional perspective)。前者指的是测量工具与测量过程的准确性(accuracy)和可推论性(generalisability)，后者则把核心放在测量目标的实现程度上。Messick(1988)认为效度是一个整体概念，要实现对测试结果的解释与运用，证据与后效缺一不可。他把构念效度升级为整体效度的代名词，一项测试的构念效度是对以上所有方面的综合考虑。需要强调的是，所谓整体效度，不是一个效度，而是对效度的方方面面进行综合判断。具体到每一项测试，达到效度的整体性是一个视具体测评目的、测评对象和测评环境而平衡的过程。

在实际操作中，构念效度面临两大威胁：无关性(irrelevance)和代表性不足(under-representation)。避免无关性内容相对容易，但代表性不足往往是一个严重问题。无论测试时间有多长，测试内容都只能覆盖基础知识或目标能力的一小部分。

语言测试任务应与什么相关并具有代表性？或者说，评估内容是什么？根据不同的目的，评估内容的界定主要基于以下三个方面：1)目标语言使用域的需求分析(target language use domain analysis)；2)课程目标(curriculum standards)概述；3)语言能力构念。例如，如果测试旨在了解以英语为媒介的大学学习所需的英语能力，则应分析高等教育领域中学习所需的典型语言任务(目标域任务需求)，且测试任务应与这个领域的任务相关并具有代表性。如果测试旨在衡量学生本学期的学习成绩(课程目标)，那么该测试应包含本学期课程提及的知识、技能和能力。如果测试与课程无关，或者包括教师未教授的内容，那将是一项不好的测试。如果测试目的与任何课程无关，甚至没有具体的未来目标语言使用域，而只想了解被试外语能力的高低，评测内容则主要依据语言能力的理论构念，即对语言能力的理论分析。结构主义语言理论下的测试包含语音、语法、词汇和听说读写能力，而强调语言功能和交际能力的理论下的测试则会考查被试语言运用任务的完成情况。

在整体构念效度的视角下，除了测试内容之外，效度的概念主要包括分数解释合理或不合理，

以及对从测试中获得的信息使用适当或不适当。针对某个目的和目标人群设计的测试可以完全与其他目的或人群无关。从这个意义上说，测试的效度不存在于测试本身，而关乎测试分数的解释和使用有效与否。打个比方，一把尺子无论用料多么精良、刻度多么准确，也不适合测量两座城市之间的距离。因此，对测试结果解释和使用的适当性是效度的重要指标。

信度指的是测试结果的稳定性，是效度的必要条件。测试信度主要受到以下三方面误差源的影响：1）同一学生在不同题目或任务上的表现不同；2）同一学生在不同时间的测试表现不同；3）不同评分员对同一学生的评分结果不同。通常采用重测法测量信度的稳定性，重测分数之间的相关系数是检验信度的指标。此外，还可以通过查看相似测试任务和题目之间的内部一致性来检验测试的信度。从理论上讲，当设计类似的测试任务和题目来考查相同能力时，受试应在这些任务中发挥相似，这被称为测试的内部一致性信度（internal consistency reliability）。检验内部一致性信度的常见方法包括：1）Cronbach's α系数，基于题目之间的相互关系；2）分半法，简单地将所有奇数项和其余偶数项构成两个单独的测试，对这两个部分进行相关检验并得到信度指标，学生的分数应在这两个部分之间保持一致。由于信度与题目数量联系紧密，将一项测试改为两项较短的测试会降低整个测试的信度。为了避免这一情况，斯皮尔曼–布朗公式（Spearman-Brown formula）通常被用于调整对半信度评估，以测量完整测试的信度（Brown 2001）。

另外，大部分教育测评是对学生表现（performance）的测评。与表现测评相关性最高的信度指标应是评分者间信度（inter-rater reliability）和评分者内信度（intra-rater reliability）。前者指不同评分者之间的一致性，后者指同一评分者在不同时间对同一个学生进行判断的一致性。

除了效度和信度之外，可行性也是需要关注的标准。可行性与资源有关，是测试设计者在决定实施测试时必须考虑的问题。理想情况下，测试应具备开发、效度验证、管理、评分和结果使用所需的所有资源。但实际上，时间、空间、财政支持和可用的人力资源可能远不及所需。一项可行的测试指的是其所需资源不超过可用资源的测试。

上述质量标准适用于大规模、高风险、标准化测试。形成性评估，特别是课堂形成性评估能否作为标准化测试的有效补充，进而改善学生学习，在很大程度上取决于形成性评估的质量。

2.2 形成性评估的效度

对于非标准化的课堂形成性评估而言，传统教育测评的质量标准是否依然适用呢？Gipps（1994）主张舍弃标准化测试奉行的具有计量学意义的信度和效度理论，并提出一套用于衡量形成性评估质量的新标准，如课程忠实度、可比性、可靠性、公信度、语境描述、公平性等。多数学者认为，效度与信度的概念依然适用。例如Shepard（2006）指出，效度在形成性评估中指的是评估结果的解释和使用能在多大程度上提高学生的语言能力，而信度指的是评估过程中判断的一致性，是效度证据的条件之一。不过，将标准化测试的效度概念应用于形成性评估，构念效度与评估结果的解释和使用含义还是有所不同的。

如果我们沿用Cronbach（1988）提出的理解效度概念的两个视角，那么对形成性评估的观察同样具有测量视角和功能视角。换句话说，形成性评估的效度也应包括评估的准确性、可推论性和评估的功能性。从测量的视角看，传统教育测评中的效度检验主要通过心理测验学（psychometrics）检测一套考题是否具有客观的相关性、代表性和分数的可推论性，并把随机误差降到最低；而形成性评估主要关心评估什么和怎么评估才能得到信得过的评估结果。从功能的视角看，传统教育测评中的效度检验把教育测评看作实现某些教育理想的工具，因此要看测试的预期效果与非预期效果；而形成性评估关心的是运用的评估手段能在多大程度上促成教学与学习的进步。

在课堂教学中，教师需要明确教学目标，了解学生现有水平与目标水平的差距，并提出缩小

差距的有效方法（文秋芳 2011）。从本质上说，形成性评估首先探究学习者现有水平与目标水平的差距。因此，首要任务就是确定目标，然后才能向目标靠近。在形成性评估中，明确的教学目标和学习目标发挥重要的导向作用，同时为确认教学成功与否提供了参照依据。任何有用的学习目标必须是一个明确的目标。目标决定了收集、解读和使用信息的方式（Cowie & Bell 1999）。教师对于目标任务和任务达成标准的认识将决定评估工具的选择、设计和使用方式。

除明确教学目标和学习目标之外，教师还应明确成功达成学习目标的标准。通过课堂语言评估，教师对课程标准、语言能力概念以及教和学的认识将指导其设计和选择工具，以提高学生的语言能力。评估什么、如何评估、如何解释评估结果、向学生提供什么样的反馈、是否及如何采取后续行动都取决于评估者对评估目标和成功标准的理解。在课堂形成性评估中，教师的评估目标主要来源于课程标准和语言能力的理论构念。教师对标准的理解和解释构成了学科教学知识（pedagogical content knowledge，简称PCK）的重要组成部分。

2.2.1 形成性评估内容

形成性评估无论采取何种形式，都必须与教学目标和学习目标相关。形成性评估任务一般是课程中指定的语言学习任务。形成性评估的内在本质表明评估任务与课堂教学和学习密切相关，要推动教、学、评一体化实施，促进学生更有效地开展学习。许多形成性评估任务实际上就是教学和学习任务。此外，形成性评估为学生提供了多种展示其理解和学习进展的机会。从这个意义上说，与传统测试相比，形成性评估的课程目标覆盖范围更广，多个练习和观察机会可以更清楚地呈现学生的真实能力和成绩。

除了评估内容的相关性和代表性外，形成性评估还强调从评估中获取信息的充分性和准确性。任何课堂活动都具有形成性评估的潜质，教师和学生应关注各种活动中可能出现的形成性评估信息，并对其进行判断和分析，用于调整教学，以提升学生的能力。一般来说，获得的信息越详细，关于相同表现的信息来源越多样化，信息就越丰富，信息的准确性也越高。如果评估信息不是对学习者成绩的准确反映，那么这些信息不仅是无用的，有时还是有害的。

2.2.2 形成性评估结果的解释和使用

Black & Wiliam（2003）认为无论采取何种评估活动，只有评估信息被教师和学生用于调整教学与学习活动的反馈时，这种评估才有利于学习。为满足学习需求，当学习证据被用于调整教学时，评估即为形成性评估。对于面向学习的形成性评估，评估信息的解释和使用至关重要，因为获取信息只是开始，除非解释评估信息并用于实现有针对性的结果，否则任何评估都不会达成形成性目标。

评估信息的解释在很大程度上取决于评估者对评估目标的理解和对学习者现有水平与评估目标差距的判断。Sadler（1989）认为，反馈只有被用于缩小学习者现有理解水平和理想理解水平的差距时，才能体现其形成性。Stiggins（2002）指出，形成性评估要求教师使用课堂评估提供的信息来促进学生学习，而不只是检查学生的学习状况，这对教师的评估素养提出了很高的要求。从课堂形成性评估任务中获得的信息是学习者的表现，且比考试分数提供的信息更多。如果对评估结果解释不当，那么教师反馈对学生学习的促进效果微弱，甚至会产生负面影响。例如，教学目标是能够在特定情况下使用某些词汇和语言结构进行投诉，而教师的现场解释若只关注语法的准确性，给予学生的反馈就会集中在语法上，导致学生的学习目标出现偏差，造成语法正确、投诉却未达到效果的局面。

在高风险、大规模的终结性评估中，解释比使用更重要，而在课堂形成性评估中，使用则比解释更重要（Nichols *et al.* 2009）。是否以及如何使用评估信息决定了评估任务的形成性。学生根据从评估任务中获得的信息采取的后续行动是构成评估形成性的关键一步。尽管如此，并不是所有的行动都有利于形成性目标的实现。例如，如果评估信息揭示出学习者的词汇问题在于其对所知单词缺乏深度掌握，学习者却被告知去记更多

生词，那么该学习者的行动也许会增加词汇量，但不太可能很快缩小词汇深度方面的差距。

形成性评估的效度来自多次反复评估，如对不同任务和不同信息来源的评估，以及给予被评估者的详细反馈和多次改进机会。就像人们看病需要反复检查各项指标，甚至到不同医院找不同医生才能确诊，而且需要进行多个疗程的治疗和再检查，形成性评估的效果也不是一次评估循环就能实现的。因此，虽然学习效果是检验形成性评估效度的最终指标，但内容效度、解释和使用效度也非常重要。换句话说，学习是一个长期的过程，不进行正确的形成性评估往往达不到应有的效果，但即便该有的步骤都有，也不一定就能实现理想效果。只要目标明确，形成性评估的步骤完整，那么就可以说这次评估具备了应有的效度。诊断正确、治疗得法，即便治不好病也不该归咎于医生。

2.2.3 形成性评估的一致性

虽然形成性评估已融入教学过程，但仍是一种评估。与标准化测试一样，形成性评估也是推断的过程，教师对学生学习情况的评估是基于观察到的学生表现作出的判断或推断（Bennett 2011）。当对学生学习的判断是基于对各种语言任务的多重观察时，判断的一致性就成为一个至关重要的问题（Parkes 2013）。课堂形成性评估的一致性显然不再是统计学问题。众所周知，人类的判断存在任意性和主观性。为了使评估信息对教学有用，教师对学生学习得出的结论需要尽量保持一致。当教师积累了更多基于课堂任务的判断经验且学生的表现更加稳定时，判断的一致性也会提高。

由于课堂形成性评估主要取决于教师对学生表现的评估，因此，最相关的信度指标应是评分者内信度和评分者间信度。多次观察和判断是课堂评估的另一特点，因此，信度的稳定性也尤为重要。对评估目标和成功标准的理解越明晰，评估的一致性就越高。同理，评估经验越丰富，就越能减少判断失误的次数。

需要注意的是，信度概念在形成性评估中有一个悖论问题。形成性评估的最终目的是使学生进步，而当学生进步明显时，就会造成评估结果前后不一致。这种情况极易出现在中长期形成性评估循环中。有时学生的课堂表现本身可能就不一致。形成性评估以学生本人为参照，即使获得同样的结果，学生得到的反馈也可能不同（Stobart 2006）。课堂评估的低风险特质使评估的低一致性变得可容忍，因为教师有机会在之后的交互活动中发现并纠正不准确的判断（Black & Wiliam 2006）。

3. 结语

形成性评估不一定总是有利于教和学，尤其当其本身也可能存在质量问题时。本文强调，明确形成性评估的概念和评估目标至关重要。虽然具体细节有所不同，但传统的语言测评质量框架依然可应用于形成性评估，特别是效度和信度等基本测试质量标准有助于理解形成性评估的质量。一套操作性强的形成性评估质量衡量标准和效度验证框架虽然是发挥形成性评估良好作用的关键，但在实践中，抽象、复杂的效度验证框架在指导教师评估实践的过程中缺乏可操作性，对灵活动态的课堂评估效度验证提出挑战。未来我们可对形成性评估效度验证方法进行探索，并展开更为深入全面的分析。

参考文献

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing* [M]. Washington: American Educational Research Association.

Bennett, R. E. 2011. Formative assessment: A critical review [J]. *Assessment in Education: Principles, Policy & Practice* 18: 5-25.

Berry, R. 2008. *Assessment for Learning* [M]. Hong Kong: Hong Kong University Press.

Berry, R. & B. Adamson (eds.). 2011. *Assessment Reform in Education: Policy and Practice* [C]. Berlin: Springer.

Black, P. & D. Wiliam. 1998. Inside the black box: Raising

standards through classroom assessment [J]. *Phi Delta Kappan* 80: 139-144, 146-148.

Black, P. & D. Wiliam. 2003. In praise of educational research: Formative assessment [J]. *British Educational Research Journal* 29: 623-637.

Black, P. & D. Wiliam. 2006. The reliability of assessments [A]. In J. Gardner (ed.). *Assessment and Learning* [C]. London: SAGE. 119-132.

Black, P. & D. Wiliam. 2009. Developing the theory of formative assessment [J]. *Educational Assessment, Evaluation and Accountability* 21: 5-31.

Bloom, B. S. 1968. Learning for mastery [J]. *Evaluation Comment (UCLA-CSEIP)* 1: 1-12.

Bloom, B. S., J. T. Hastings & G. F. Madaus (eds.). 1971. *Handbook on Formative and Summative Evaluation of Student Learning* [C]. New York: McGraw-Hill.

Broadfoot, P. & P. Black. 2004. Redefining assessment? The first ten years of assessment in education [J]. *Assessment in Education: Principles, Policy & Practice* 11: 7-26.

Brookhart, S. M. 2003. Developing measurement theory for classroom assessment purposes and uses [J]. *Educational Measurement: Issues and Practice* 22: 5-12.

Brown, J. D. 2001. Can we use the Spearman-Brown prophecy formula to defend low reliability? [J] *Shiken: JALT Testing & Evaluation SIG Newsletter* 4: 7-11.

Cizek, G. J. 2009. Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts [J]. *Theory into Practice* 48: 63-71.

Cowie, B. & B. Bell. 1999. A model of formative assessment in science education [J]. *Assessment in Education: Principles, Policy & Practice* 6: 101-116.

Cronbach, L. J. 1963. Course improvement through evaluation [J]. *Teachers College Record* 64: 672-683.

Cronbach, L. J. 1988. Five perspectives on the validity argument [A]. In H. Wainer & H. I. Braun (eds.). *Test Validity* [C]. Hillsdale: Lawrence Erlbaum Associates. 3-17.

Earl, L. M. 2003. *Assessment as Learning: Using Classroom Assessment to Maximize Student Learning* [M]. Thousand Oaks: Corwin.

Gipps, C. V. 1994. *Beyond Testing: Towards a Theory of Educational Assessment* [M]. London: Falmer.

Gu, P. Y. 2020. *Classroom-based Formative Assessment* [M]. Beijing: Foreign Language Teaching and Research Press.

Lado, R. 1961. *Language Testing: The Construction and Use of Foreign Language Tests: A Teacher's Book* [M]. London: Longmans.

Leung, C. & B. Mohan. 2004. Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse [J]. *Language Testing* 21: 335-359.

Messick, S. 1988. Validity [A]. In R. L. Linn (ed.). *Educational Measurement* (3rd Ed.) [C]. New York: Macmillan. 13-103.

Newton, P. E. & J.-A. Baird. 2016. The great validity debate [J]. *Assessment in Education: Principles, Policy & Practice* 23: 173-177.

Nichols, P. D., J. L. Meyers & K. S. Burling. 2009. A framework for evaluating and planning assessments intended to improve student achievement [J]. *Educational Measurement: Issues and Practice* 28: 14-23.

Parkes, J. 2013. Reliability in classroom assessment [A]. In J. H. McMillan (ed.). *SAGE Handbook of Research on Classroom Assessment* [C]. Los Angeles: SAGE. 107-124.

Sadler, D. R. 1989. Formative assessment and the design of instructional systems [J]. *Instructional Science* 18: 119-144.

Scriven, M. 1967. The methodology of evaluation [A]. In R. Tyler, R. Gagne & M. Scriven (eds.). *Perspectives of Curriculum Evaluation* [C]. Chicago: Rand McNally. 39-83.

Shepard, L. A. 2006. Classroom assessment [A]. In R. L. Brennan (ed.). *Educational Measurement* (4th Ed.) [C]. Westport: Praeger. 624-646.

Shepard, L. A. 2013. Foreword [A]. In J. H. McMillan (ed.). *SAGE Handbook of Research on Classroom Assessment* [C]. London: SAGE. XIX-XXII.

Stiggins, R. J. 2002. Assessment crisis: The absence of assessment for learning [J]. *Phi Delta Kappan* 83: 758-765.

Stiggins, R. J. 2005. *Student-involved Assessment for Learning* (4th Ed.) [M]. Upper Saddle River: Merrill Prentice Hall.

Stobart, G. 2006. The validity of formative assessment [A]. In J. Gardner (ed.). *Assessment and Learning* [C]. London: SAGE. 133-146.

Stufflebeam, D. L. 1983. The CIPP model for program evaluation [A]. In G. F. Madaus, M. S. Scriven & D. L. Stufflebeam (eds.). *Evaluation Models: Viewpoints on Educational and Human Services Evaluation* [C].

Boston: Kluwer-Nijhoff Publishing. 117-141.
毕鹏晖，2017，大学英语微移动词汇学习融入形成性评估模式的研究[J]，《外语电化教学》(1)：35-42。
曹荣平、陈亚平，2013，形成性评估及其在口译教学中的应用探析[J]，《中国翻译》(1)：45-50。
范劲松、季佩英，2017，翻译教学中的师评、自评和互评研究——基于多层面Rasch模型的方法[J]，《外语界》(4)：61-70。
顾永琦、余国兴，2018，如何研究课堂形成性评估——以“中国基础教育外语测评研究基金”立项课题为例[J]，《英语学习(教师版)》(10)：39-44。
黄剑、罗少茜、林敦来，2019，国内外语教育形成性评价研究述评：回顾与建议[J]，《外语测试与教学》(3)：1-9。
李传益，2015，非英语专业学生英语口语能力课堂即时评价实证研究[J]，《外语测试与教学》(3)：44-52。
李清华、孔文，2015，外语形成性评估的效度验证框架[J]，《外语教学理论与实践》(1)：24-31。
李雪莲，2016，促进学习的课堂评价及学习目标的自我管理研究[J]，《现代外语》(3)：399-407。
刘建达，2019，中国英语能力等级量表与英语教学[J]，《外语界》(3)：7-14。
卢耀增，1987，教育评价应突出形成性评价[J]，《天津教育》(5)：6-7。
罗少茜、张帅，2019，运用学习档案袋评价发展学生英语学科核心素养[J]，《外语测试与教学》(2)：33-40。
潘鸣威、吴雪峰，2019，中国英语能力等级量表在中小学英语形成性评价中的应用——以写作能力为例[J]，《外语界》(1)：89-96。
唐雄英，2017，评价课程改革与中小学英语教师评价素养发展[J]，《外语测试与教学》(1)：21-29。
田朝霞，2018，“英语演讲”课在中国高校的本土化再研讨——形成性评估课程设计与实施[J]，《外语教育研究前沿》(1)：26-34。
王慧文、刘芹，2018，理工科大学生学术英语听说课程电子档案袋评价模式探索[J]，《外语测试与教学》(1)：1-11。
文秋芳，2011，《文献阅读与评价》课程的形成性评估：理论与实践[J]，《外语测试与教学》(3)：39-49。
文秋芳，2016，“师生合作评价”：“产出导向法”创设的新评价形式[J]，《外语界》(5)：37-43。
夏晓燕、林敦来、郭乙瑶，2019，非英语专业研究生通用学术英语听力校本测评体系的开发[J]，《外语教育研究前沿》(1)：66-72。
杨华、文秋芳，2014，外语课堂即时形成性评估的“相倚性”研究[J]，《外语教学》(4)：41-45。
袁树厚、束定芳，2017，我国外语教学中的形成性评价研究：回顾与思考(2002—2016)[J]，《外语教学理论与实践》(4)：51-56。
曾永红、梁玥，2017，不同类型同伴互评对大学生英语写作的影响实证研究[J]，《外语研究》(4)：53-57。
张荔，2017，学术英语交际课程形成性评估模式及效果研究[J]，《中国外语》(2)：72-80。
中野照海、陈穗申、袁洪国，1989，视听媒体开发的理论[J]，《外语电化教学》(4)：43-44。

作者简介

顾永琦，新西兰惠灵顿维多利亚大学语言学与应用语言研究学院副教授，博士生导师。主要研究领域：语言测评、学习策略、词汇教学、教师发展。电子邮箱：peter.gu@vuw.ac.nz

李加义，新西兰惠灵顿维多利亚大学语言学与应用语言研究学院博士研究生。主要研究领域：语言测评、外语教学。电子邮箱：jiayi.li@vuw.ac.nz

（审稿编辑：许宏晨）