导　　读

《第二语言口语测评研究与实践》（以下简称《口语测评》）（*Examining Speaking: Research and Practice in Assessing Second Language Speaking*）是英国剑桥大学出版社“语言测试研究”（Studies in Language Testing，简称 SiLT）系列丛书的第 30 册，由剑桥考试中心顾问 Lynda Taylor 主编。该书以 Weir（2005）提出的社会—认知效度框架（socio-cognitive validity framework）为理论基础，从不同的方面论证了剑桥口语测试的效度。本文对该书的内容作简要介绍和评述，以便读者更好地理解该书所推介的理论框架，并在今后的口语测试开发和效度研究中有效地运用该框架。

一、理论框架

现代语言测试对二语学习者口语能力的评价大多采用行为测试（performance assessment）。McNamara（1996）提出了一个行为测试特征模型。该模型由五个要素组成：考生、工具、考生在测试中的表现、评分员和评分量表。其中，考生表现是该模型的核心要素，测试开发和效度研究都围绕着考生表现展开。考生和工具是决定考生表现的重要因素。考试设计者根据考生特征和测试目标，完成测试工具的设计和开发，并通过各种施考手段（如面试等），在尽可能真实的测试环境下采集考生口头表达的样本，全面评价口语能力构念的各个维度。评分员和评分量表是影响测试效度的关键因素。面试型测试中，考官的

角色是主考，即通过提问或对话等形式主持考试；考官的任务是评分，即运用恰当的评分量表，在考试过程中对考生表现进行主观评判。非面试型口语测试（如计算机化考试）采用预先录制的音频或视频替代主考，评分过程与考试过程分开，一般在考后完成（Luoma 2004）。

为了更好地指导考试开发和效度研究，Weir（2005）提出了社会—认知效度框架。该框架与 McNamara（1996）的行为测试特征模型相似，以考生的答题反应为核心，从围绕考生表现的前期效度和围绕分数解释和考试使用的后期效度对考试效度进行全面论证。其中，前期效度的三个主要维度是考生特征、认知效度和语境效度；后期效度的三个关键维度是评分效度、后果效度和校标关联效度。该框架既考虑了考生答题的认知过程，又兼顾了考试任务设计的语境特征和考试使用的社会性，为考试开发、考试效度证据的采集和论证提供了细致且可操作的理论框架。

二、编排结构

SiLT 系列丛书中共有四部作品运用 Weir（2005）的社会—认知效度框架，分别探讨了写作测评（Shaw & Weir 2007）、阅读测评（Khalifa & Weir 2009）、口语测评（Taylor 2011）以及听力测评（Geranpayeh & Taylor 2013）的开发和效度研究。这四部作品被称为 SiLT 系列丛书的“构念”分册。

《口语测评》一书围绕剑桥口语测试（KET, PET, FCE, CAE, CPE）的开发和效度研究展开了深入的探讨。全书共八个章节。第一章介绍了全书的理论基础、编撰意图和主要内容。第二至第七章基于社会—认知效度框架，以剑桥口语测试为例，从以下六个方面回顾和总结了口语测试的开发和效度论证：考生特征（第二章）、认知效度（第三

章）、语境效度（第四章）、评分效度（第五章）、后果效度（第六章）以及校标关联效度（第七章）。第八章是全书总结。以下为该书各章内容的简要介绍。

三、内容简介

第一章由主编 Lynda Taylor 撰写。首先，作者简要介绍了该书编撰的理论基础，即 Weir（2005）的社会—认知效度框架，并强调该框架的意义在于指导我们对测试的理论和实践进行系统的、全面的检验和论证，为考试效度提供理论依据和实证数据。其次，作者指出该书适合的读者群体和编撰的意图。该书最直接相关的读者群体是从事剑桥口语测试开发、实施和培训的专业人员；该书也适合参与其他口语测试开发和研究的相关人员，包括考试机构和教育决策机构的人员、测试研究者、测试专业的学生等；该书还有助于提高担任口语教学的一线教师、口语教学大纲设计和教材开发者、招生工作者等对剑桥口语测试的理解和认识，以利于他们更好地使用剑桥口语测试。总之，作者认为对效度的探索和研究有助于构建一个开放、透明、有责任感的语言测试道德氛围，这样的道德氛围将有助于研究和改进考试的社会后果和教学后效。作者在该章中对剑桥口语测试的能力等级、考试系列以及传统特色也做了详细的介绍，为后面的章节做好了扎实的铺垫，帮助读者更好地理解剑桥口语测试的开发和效度研究。

第二章的作者是 Barry O’Sullivan 和 Anthony Green。在社会—认知效度框架中，考生的答题反应是测试的核心要素，考试设计的目标是使考生的答题表现尽可能真实，即答题的认知过程与真实语言交际的认知过程相符（见第三章）。为实现这一目标，考试设计者需要清楚地了解影响答题过程的所有考生相关特征。为此，作者基于 Brown

（1995）的研究和其他前期研究，提出了一个考生特征框架。该框架把考生特征分为三大类：物理 / 生理特征，即考生年龄、性别、身体条件等；心理特征，即考生的个性、记忆力、认知特点、情感、注意力、动机、情绪等；经验特征，主要包括考生的教育和学习经历、文化意识、交流沟通能力等。考生特征还可以分为群体特征（如年龄）和个体特征（如残疾状况）。作者强调，不同类型的考生特征并非独立地对考试产生影响，而是交互作用于考生在考试中的行为表现。基于该分类框架，作者分析了剑桥口语测试中的各种考生特征，展示了如何全面、系统地采集考生特征的相关信息，研究考生特征对考试产生的影响，并解决由此可能产生的考试公平性问题。

第三章由 John Field 撰写。认知效度属于考试的前期效度，关注的是考生答题过程，而不是考试结果，因此需要在考试设计和开发阶段进行论证。作者指出，认知效度研究的首要问题是测试活动在多大程度上体现真实的语言运用；其次，认知效度研究应探索不同口语水平的考生的答题认知过程有何不同，即考生答题的认知过程能否区分不同语言水平的考生。基于 Levelt（1989）的研究以及对口语能力本质的理解，作者提出了一个由六个阶段组成的口头表达认知效度框架：概念化、语法编码、音位编码、语音编码、发音、自我监控。作者细致地阐述了各个认知阶段的特点，并运用该框架分析了剑桥口语测试所体现的认知效度。作者分析了各项考试的考试细则，描述了考试在保证认知效度方面的各种实践举措。这些措施既涉及显性的认知维度，如任务设计（考生对话题的熟悉程度、测试任务所要求的交际功能等）、考生表现（犹豫、停顿、语块运用、语音清晰度等），也涉及隐性的认知过程，如组句成篇的过程、自我监控策略的运用等。作者还说明了这些认知维度在剑桥五个级别口语测试中的差异。此外，作者特别分析了影响口语测试任务难度的两个重要认知特征：1）口语表达

中的时间压力（主要是准备时间）；2）交互性口头交际中的对话者权力、语篇共建、听力理解等问题。作者指出，认知效度的实现主要通过考试的设计和实施，而考试的设计和实施也决定了考试的语境效度。因此，认知和语境是考试效度不可分割的两个重要组成部分。

第四章由 Evelina Galaczi 和 Angela ffrench 撰写。语境效度与认知效度一样，属于考试的前期效度，其核心是影响考生表现的各种语境特征。作者根据社会—认知效度框架，将口语测试的语境参数分为三大类：任务设计，如答题模式、测试目的、权重分配、任务顺序、准备和答题时间等；测试实施，即施考环境、实施流程、考试安全等；任务要求，包括与语言能力相关的要求，如交流模式、话语模式、话题熟悉程度等，还包括与对话者特征相关的要求，如语速、口音、参与人数、性别等。作者对这些语境特征进行了逐一解释（限于篇幅，测试实施的语境分析在附录 D 中呈现），并分析了这些特征在剑桥口语测试开发和效度研究中的运用。作者对语境效度的分析从两个层面展开：一是语境特征对任务设计、测试实施和任务要求的总体影响，二是语境特征对不同级别考试产生的不同影响。作者以任务设计中的答题模式特征为例做了进一步阐述。首先，作者汇总了剑桥五个级别口语测试所采用的答题模式（表 4.1, pp. 133–134），并以交际语言能力理论为依据，分析了这些答题模式对考试构念的体现程度；其次，作者对比分析了剑桥五个级别口语测试中答题模式的开放程度，说明了对该语境特征的调控可以影响口语测试的等级难度（图 4.2, p. 137）。

第五章的作者是 Lynda Taylor 和 Evelina Galaczi。评分效度关乎考试的信度和分数的解释。在口语测试中，评分效度的影响因素包括评分标准和评分量表、评分过程、评分条件、评分员特征、评分员培训、分数监控、等级评定和分数报告等。作者逐一阐述了这些影响评分的因素并分享了剑桥口语测试的经验。首先，作者分析了整体评分和分

项评分的利弊，并介绍了剑桥口语测试的评分标准和量表。剑桥口语测试基本采用面试形式，包括一对一面试形式或两名考生和两名考官的小组面试形式。小组面试中一名考官（通常是主考）采用整体评分，另一名考官则采用分项评分；评分标准包括语法和词汇、语音、语篇能力、交互能力等。作者指出，小组形式的评分难点在于话语共建，即评价的究竟是某考生的表现还是小组的表现。其次，作者分析了评分过程（如评分策略、信念、行为）、评分条件（如面试的场地、环境、评分时间）和评分员特征（如性别、年龄、语言和文化背景、评分严厉度）对评分效度的影响，阐述了评分员培训的意义、方法和效果。作者指出，一味追求评分一致性可能产生负面影响，即造成评分员失去独立、主观判断的意愿，而仅仅对表面化的标准（如语音、语调）做简单的评判。最后，作者介绍了口语测试的等级评定和分数报告，指出要特别关注位于两个级别分界线附近的非典型考生的等级评判。

第六章由 Roger Hawkey 撰写。作者首先界定了后果效度的两个核心概念：影响和后效。前者指考试的使用对社会和机构所产生的影响；后者指考试的实施对课堂教学或工作环境中的个体所产生的影响。作者强调了考试后果的复杂性，并从 Messick（1996）提出的构念代表性不充分和构念不相关两个方面，阐述了考试后果与考试效度（语境、认知和评分效度）的关系，即两者之间并非直接的、简单的对应关系，而是受许多复杂的环境因素（如教师的语言水平、培训经历、动机和动力、课程长短、班级人数）的影响。作者指出，后果效度研究必须基于证据，且必须证明考试后果与考试的实施和使用相关。而且，后果效度研究应该贯穿于考试开发和使用的每个阶段，包括考试对课程标准、课程设计、教材、备考、施考、评分等的影响，也包括考试决策和使用所产生的各种后果。高风险考试的后果效度还应涵盖考试的

伦理道德和公平公正性。此外，考试后果是双向的，即考试对考生、教师和其他利益相关者产生影响；同时，这些后果反过来影响考试，对考试的进一步改革和发展至关重要。作者以雅思口试为例，分析了后果效度研究的范畴，并介绍了其他剑桥口语测试的后果效度研究，进一步说明了后果效度与考试效度的其他各个维度密不可分。

第七章由 Hanan Khalifa 和 Angeliki Salamoura 撰写。校标关联效度的证据来自三个方面：不同考试之间的关联、同一考试的不同考次之间的关联、考试与外部校标的关联。不同考试之间的关联通过分数等值实现，前提是两者的测试目的和测试群体基本一致，并且考核基本相同的能力；等值一般采用 IRT 或回归方法，也可以通过对接量表进行。考试之间的关联分为同期效度和预测效度，前者是同一批考生在相近的时间内参加两个考试，后者是同一批考生在间隔较长的时间内参加两个考试，前一次考试成绩预测后一次成绩。作者以剑桥 FCE 考试与 ETS 的 SPEAK 考试对比研究为例，介绍了如何通过考试之间等值开展校标关联效度研究。第二种校标关联效度的证据，即同一考试的不同考次之间的关联，是建立考次之间的等值，即同一批考生在相同的条件下参加不同批次的考试。等值方式可以采用定量分析（如难度、相关、ANOVA、MFRM）；对于口语测试来说，更重要的是定性分析（如测试内容分析、考官表现分析等），以确保不同批次的考试在语境和认知的各个参数上具有可比性。最后，作者介绍了与外部标准关联的校标效度。外部标准是指具有一定影响力的语言能力量表或评价体系，其中欧洲语言共同参考框架是最重要的标准之一。作者以 FCE 与 CEFR 的对接为例，展示了外部标准关联的方法。作者还提出了考试与量表对接的局限性：不同的考试目的不同，适用环境不同，与量表对接并不意味着这些考试可以互相替代。

在第八章，Cyril Weir 和 Lynda Taylor 共同为该书做了全面的总结，

并进一步说明了社会—认知效度框架对考试开发和效度研究的意义。作者指出，社会—认知效度框架指导下的语言测试效度观既体现了语言交际的社会属性，也反映了语言交际中复杂的认知过程。该框架是对剑桥考试中心 VRIP（Validity, Reliability, Impact and Practicality）框架的改进和提升，能够更加透彻地阐述考试效度的各个维度以及这些维度之间的关系，为考试开发和效度研究提供更好的指导作用。在社会—认知效度框架指导下，考试效度研究需要分析语言交际任务的语境特征，探索考生答题和考官评分的认知过程，关注考试对社会、机构和个人的影响；更为重要的是，认知、语境和评分等维度对考试产生交互影响，共同决定着考试的效度。

四、评述和结语

《口语测评》一书基于 Weir（2005）提出的社会—认知效度框架，从考生特征、认知效度、语境效度、评分效度、后果效度和校标关联效度等方面探索了剑桥口语测试的效度。社会—认知效度框架的六个维度既互相独立又互相关联，其中认知、语境和评分三个方面构成了该框架的“核心三角”，对实现考试的理论构念至关重要（Taylor & Galaczi 2011）。值得一提的是，该书与其姊妹篇《阅读测评》和《写作测评》不同，是一部由不同作者完成的编辑著作，作者是来自英国多所大学的语言测试研究者，剑桥考试机构的开发者、研究者和顾问。正因如此，该书从不同的视角，全面、系统地阐述了剑桥口语测试的效度，对口语测试的开发者和使用者具有重要的启示，也为口语测试的效度研究提供了切实可行的思路和方法。

效度研究是语言测试领域一个经久不衰的主题。经过半个多世纪的发展，语言测试研究者对效度的理解不断加深，同时也摸索出了

一些效度论证的理论模式。其中基于论证的模式（an argument-based approach）在世界各地，特别是北美地区，影响最为广泛。Bachman（2005）和 Bachman & Palmer（2010）指出，效度论证包括两个递进的环节：1）测试效度论证，探索考生表现与分数之间的关系；2）测试使用论证，重点关注测试使用及其产生的后果。第一个环节关注的就是社会—认知效度框架中的“核心三角”问题，即考试构念的定义及其在考试中的实现；第二个环节关注了社会—认知效度框架中的“后果效度”。因此，源于欧洲的社会—认知效度框架与始于北美的论证模式对效度的理解和定义都基于整体效度观（Messick 1989），以构念效度为核心，重视考试所带来的价值以及考试使用所产生的后果。无论何种论证模式，效度研究都强调基于多方面的证据，让事实说话。相对而言，社会—认知效度框架更加明确地关注考试任务的语境特征和考生答题的认知过程，并且倡导通过考试设计改进考试后效。考试一旦开发成功并被投入使用，就有了属于自己的、不受开发者所掌控的生命周期，因此考试设计是确保考试产生良好后果的重要基础。

当然，从语言测试近期的发展趋势来看，该书仍有一些不足之处。首先，SiLT 系列丛书的构念分册以听、说、读、写四大语言技能为基础，探索语言测试的构念和语言测试的效度。但是，语言测试已经开始走向语言技能综合的发展趋势（Taylor 2011）。作为案例分析的剑桥口语测试主要为传统的单技能测试，因此，该书基本未涉及“听—说”“听—读—说”等技能综合的测试任务。对此类口语测试的构念界定、任务设计、评分标准、分数解释等都有待进一步探索和研究。其次，由于剑桥口语测试主要采用面试，因此该书未充分讨论计算机化口语测试的效度论证（Chi 2013）。随着网络和通信技术的快速发展，大规模考试逐步向计算机化考试转型。目前，托福网考和培生学术英语考试的口试都采用机考；我国大学英语四、六级口语考试从 2015 年

起全面实施机考，而且采用双人小组的形式，即两名考生共同完成朗读、问答、陈述、讨论等测试任务。计算机化口语测试与面试有很多不同之处，涉及社会—认知效度框架中的各个维度，如考试任务设计、试题材料的呈现、主考的方式、考生答题模式、评分方式等。因此，我们需要进一步探索如何运用社会—认知效度框架，针对计算机化口语测试的特点，全面论证口语测试的效度。

金艳

2019 年 8 月 30 日

参考文献

Bachman, L. F. (2005) Building and supporting a case for test use. *Language Assessment Quarterly*, 2 (1), 1–34.

Bachman, L. F. & Palmer, A. S. (2010) *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World.* Oxford: Oxford University Press.

Brown, A. (1995) The effect of rater variables in the development of an occupation-specific language performance test, *Language Testing* 12, 1–15.

Chi, Y. (2013) Book review. *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing 30. Edited by Lynda Taylor, *Language Assessment Quarterly, 10* (4), 476–479.

Geranpayeh, A., & Taylor, L. (Eds) (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening,* Studies in Language Testing 35. Cambridge: Cambridge University Press.

Khalifa, H. & Weir, C. J. (2009) *Examining Reading: Research and Practice*

in Assessing Second Language Reading, Studies in Language Testing 29. Cambridge: UCLES/Cambridge University Press.

Levelt, W. J. M. (1989) *Speaking*. Cambridge, MA: MIT Press.

Luoma, S. (2004) *Assessing Speaking*. Cambridge: Cambridge University Press.

McNamara, T. F. (1996) *Measuring Second Language Performance.* London: Longman.

Messick, S. (1996) Validity and washback in language testing, *Language Testing* 13(4), 241–256.

Messick, S. A. (1989) Validity, in Linn, R. L. (Ed), *Educational Measurement* (3rd edition). Washington DC: The American Council on Education and the National Council on Measurement in Education, 13–103.

Shaw, S. D. & Weir, C. J. (2007) *Examining Writing: Research and Practice in Assessing Second Language Writing*, Studies in Language Testing 26. Cambridge: UCLES/Cambridge University Press.

Taylor, L. (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing 30. Cambridge: UCLES/Cambridge University Press.

Taylor, L. & Galaczi, E. (2011) Scoring validity, in Taylor, L. (Ed.) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing 30. Cambridge: UCLES/Cambridge University Press. 171–233.

Weir, C. J. (2005) *Language Testing and Validation: An Evidence-based Approach.* Basingstoke: Palgrave Macmillan.