导　　读

《第二语言写作测评研究与实践》一书由 Stuart D. Shaw 和 Cyril J. Weir 两位学者合作完成，并由剑桥大学出版社于 2007 年出版。作为剑桥大学语言测试系列丛书中的一部专著（系列编号 26），该书以 Weir（2005）所提出的社会—认知模型 / 框架（socio-cognitive model/framework）为主要研究框架，系统地从多维角度对二语写作测评的效度加以举证，并梳理了写作测评中不同效度维度所采用的效验方法。特别值得指出的是，作者以剑桥系列英语考试（Cambridge Main Suite examinations）的写作测试为主要举例对象，给读者呈现了该系列考试中写作测试效度的各类证据和种种尝试。

该书出版后，语言测试学界中专门研究写作测试的专家 Liz Hamp-Lyons（2011）和 Sara R. Weigle（2010）曾分别发表过书评。前者是写作研究权威期刊 *Assessing Writing* 的创刊主编，后者是剑桥大学出版社另一套测评丛书之一 *Assessing Writing*（Weigle 2002，外语教学与研究出版社于 2011 年已引进）的作者。两篇书评均从不同角度针对本书的内容、意义、突破性和局限性等加以评论。读者也可在阅读完本书后继续查阅这两篇书评。

一、本书概述

语言测试研究发展到 21 世纪，研究者的目光越来越聚焦高利害考试在设计开发、效度验证、施考维护等环节的透明度和公开度。因此，国际上知名的英语水平考试机构纷纷出版或公开有关考试的效度验证报告，并将此作为首要任务不断滚动发布。比如，美国教育考

试服务中心（Educational Testing Service）曾于2008年出版了新托福考试（TOEFL iBT）的效度研究专著（Chapelle, Enright and Jamieson 2008），从效度证据（validity argument）角度说明了考试本身在效度验证方面的尝试和努力。相类似的是，本书正是剑桥大学英语考试委员会（Cambridge ESOL）的一项尝试。两位作者从剑桥系列英语考试的写作测试入手，厘清了在社会—认知框架下写作测试的构念，并结合该系列考试阐述了有关写作效度验证的研究和方法。本书的系统性很强，从写作测试效度的各个方面加以分解并合成，也有机地梳理了二语写作的研究方法。

首先，本书的系统性强，为大规模、高利害考试的效度验证提供了很有价值的框架。全书共分为8章。第1章是引言，提纲挈领地介绍了本书的研究框架和整体基调。第2章是考生特征（test-taker characteristics），详细说明了写作测试中考生特征的不同方面对效度的影响。第3章是认知效度（cognitive validity），从写作过程的角度梳理了写作测试的效度举证。第4章是环境效度[1]（context validity），从客观条件、写作任务的设计等角度分析了影响写作测试效度的因素。第5章是评分效度（scoring validity），从写作评分量表和评分员等方面说明了写作测试的效度元素。第6章是后效效度（consequential validity），从反拨效应的角度说明了写作测评的效度问题。第7章为对标效度（criterion-related validity），从测量标准的参考等视角厘清了写作测试效度需要注意的方方面面。最后第8章则总结了本书的主要内容并对写作测试的开发和效度验证提出了具有建设性的意见。为更好地帮助读者理解本书中大量的实例，本书还有附录，呈现了剑桥系列英语考试的各个细节。就篇幅而言，作者在第4章和第5章所用的笔墨最多，

1 有关文献中将context validity译为“环境效度”、“语境效度”等。本书中主要指包含写作测试的任务设计和写作施考的环境及保障等，故在本导读中统一为“环境效度”。具体内容详见本书第4章。

亦是作者重点阐释的内容。读者需要对剑桥系列英语考试中写作测评的有关情况有较为全面的了解，方可更好地理解本书内容。

再者，全书结构清晰，章节之间构成一个有机的整体。第 1 章名为引言，实则详细介绍了本书的研究框架，并将 Weir（2005）的社会—认知模型应用于写作测评，给读者呈现了写作测评中该模型的具化内容，提出了从考生特征、认知效度、环境效度、评分效度、后效效度以及对标效度等六个方面对写作效度加以举证。因此，本书第 1 章是总领，后六章则为具体展开，并在最后一章总结。本书主干部分的各章节编排基本一致。作者首先从文献回顾的角度阐述了各个效度组成的有关研究，然后介绍了剑桥系列英语考试在这些效度举证方面所进行的研究，并在后记（postscript）中以问题清单的方式总结了剑桥系列英语考试在这些效度验证方面的举措和展望。

最后，本书的实践性强。作者在各个章节中对通过不同的研究方法进行的写作测试效度验证展开论述。这些方法本身也可直接应用到其他类似的英语写作考试之中。

因此，本书对我国应用语言学研究、特别是语言测评方面的研究意义可见一斑。至今为止，我们大规模、高利害的考试中仅有大学英语四、六级考试（见杨惠中 & Weir 1998）和英语专业四、八级考试（见邹申，1998）出版过相关的效度验证报告。当时的效度研究为这些考试的公平性和透明度作出了积极的贡献。然而，囿于当时的效度观和效度验证研究框架的滞后，二十余年前的效度研究已亟需更新，需从更多的视角、维度、渠道等为考试效度加以举证。此外，就写作而言，我国学者的效度研究往往点面零散，不成系统，更无研究专注于某一大规模、高利害考试的写作测试。

本书的研究范式给广大长期从事写作测评的研究人员提供了范本。我们可以依据不同的效度组成，以社会—认知模型为研究框架，系统

并长期地开展有关英语考试中写作测试的效度研究。此外，随着《中国英语能力等级量表》的问世，我国学者可有意识地增加效度研究的渠道，以多维度、多视角的方式对写作测试的效度加以举证。

二、内容导读

第 1 章　引言　第 1 章为全书的研究框架奠定了主要的基调，详细展现了 Weir（2005）的社会—认知框架，并解释了语言测试中效度观的演变。值得指出的是，作者在这一章节中把这一研究框架应用于写作测试，提出了与写作测试密不可分的几个效度组成（见本书第 4 页），即考生特征、环境效度、认知效度、评分效度、后效效度以及对标效度。这些效度组成也是本书第 2 章至第 7 章的主要内容。

考生特征主要是指考生本身在生理、心理、经验等方面的因素可能对写作测试效度所产生的影响。环境效度则是指社会文化环境、写作试题设计以及其他客观因素（如考试时间、物理环境、作答长度等）对写作测试效度可能产生的影响。而认知效度则与写作的过程有关，是检验写作过程与写作试题所触发写作思维的相符程度。认知效度在举证中可以是考前举证，如收集试测考生的口头汇报等，亦可是考后举证，如对考生成绩进行统计分析等。评分效度主要围绕评分环节展开，聚焦评分是否紧扣评分标准，以及评分员的个体差异等。后效效度属于考试反拨效应（washback）的一个方面，在本书中主要指写作测试的成绩使用以及写作测试对教和学的影响。对标效度根据所对应标准的不同，又可分为共时效度（concurrent validity）和预测效度（predictive validity）。两者均以外部标准（external indicator）为参照，分别检测写作测试与其他考试在测量结果方面的契合度和预测度。这些效度组成相互影响，并可从多个视角为写作测试提供效度证据（读者可参考图 1.1 的介绍）。

大体而言，这些效度组成以及相关证据的收集仍以 Weir（2005）

的框架为基础，前三个方面基本属于事先（a priori）效度，而后三个方面则为事后（a posteriori）效度，前后以考生完成写作作答为分界点，即考前和考后。当然，部分效度组成，如认知效度，在举证过程中也存在考前和考后共存的情况。

此外，第 1 章还全面介绍了剑桥英语考试系列的具体情况（读者可直接参照表 1.1 和表 1.2），并着重说明了这一系列考试中写作测试的能力目标及其与《欧洲语言共同参考框架》（Common European Framework of Reference，以下简称 CEFR）的对接情况。

第 2 章　考生特征　第 2 章聚焦了考生特征。作者在第 1 章研究框架的基础上回顾了考生特征的定义和分类方法。以托福考试在这方面的研究为切入点，作者指出，文献研究表明，考生群体差异，如性别、社会背景等都会对写作测试效度带来一定的影响，严重的还会对考试本身造成统计学意义上的项目偏颇（differential item functioning）。随后，作者在 O’Sullivan（2000）研究的基础上，提炼出考生特征可能会对写作测试效度造成影响的几个方面，即生理因素（physical/physiological ）、心理因素（psychological）以及经验因素（experiential）。这些方面的研究对写作测试的效度意义重大，因为这直接涉及到考试的公平性问题。设想如果考生由于生理缺陷或心理因素等而未被测量出真实的写作水平，则这一考试并没有发挥应有的作用。同理，如果有考生因为对话题熟悉或通过突击备考等方式可以取得高分，那么该项考试本身也存在问题。

鉴于此，作者汇报了一些案例研究。比如，剑桥大学考试委员会曾对生理有缺陷的考生做过一些特殊安排，对读写障碍（dyslexia）人士延长写作测试的作答时间，对盲人和弱视人士采用盲卷或字体放大的试卷等措施，以此保证考试的公平性，体现对考生的包容性。但是，在对读写障碍人士的写作作答是否采用单独评分这一问题上，有关实

证研究也发现单独评分和与正常考生作答一起评分并不存在显著差异，这对该考试的评分决策起到一定的作用。又如，在考生对考试的熟悉程度上，剑桥大学考试委员会的有关研究表明，考生对计算机操作的熟悉程度并不会对其考试压力或焦虑产生显著差异。这在一定程度上保证了机考的信度。这些举措均是从考生生理、心理和经验等因素出发来推行的，对考生特征这一效度组成提供了较多的举证。

第 3 章　认知效度　第 3 章围绕认知效度展开。作者一开始就对认知效度加以定义，即“写作测试的认知效度是检测试题与其在写作情景中所涉及的认知过程，即真实生活中完成写作任务的认知过程的契合度”（见本书第 39 页）。在这一定义下，作者除了借鉴 Weir（2005）的社会—认知框架外，也参考了 Hayes & Flower（1980）和 Field（2004, 2005）等有关认知研究的成果，指出写作测试的认知效度可以从三个写作阶段进行效度举证，即构思（planning）、转化（translating）和检查（reviewing）。

写前阶段的构思则又可细分为宏观构思（macro-planning）、组织（organisation）和微观构思（micro-planning）三个方面。宏观构思主要指思想设计以及写作障碍层面的内容，如文体、读者群、目标等。组织则对宏观构思的结果加以梳理，如分清主要论点和次要论点。微观构思则涉及如何将这些内容在段落之间和段落内较好地呈现。写中阶段的转化就是将大脑中的构思结果转换成语言文字。写后阶段的检查又分为监控（monitoring）和修改（revising）。监控在基础层面是指对拼写、标点和语法正确等的监控，而在高级层面则是指对写作意图和谋篇布局等的监控。修改是在监控的指引下对字、词、句、篇以及意义层面的改进和完善。

作者在本章中也提出了二语写作的技能问题，并通过比较写作新手和经验丰富的作者，指出前者在写作过程中往往会发生知识告知

（knowledge telling）的过程，而后者则更多采取知识转化（knowledge transforming）的策略。

由于认知过程不易观测，认知效度的举证往往也较为困难，因此，作者结合剑桥系列英语考试，以研读考试大纲（exam syllabus）和考评报告（exam report），即考官就考题在真实生活中的表现所做的点评等方式，论述了剑桥英语系列考试中写作测试的认知效度。前一种方法属于文献分析（document analysis），从考试大纲对写作过程的认知要求中提炼出写作测试的效度证据；后一种方法则属于专家判断（expert judgment），是让考官来汇报写作任务与真实生活中写作活动的匹配度。对这方面内容感兴趣的读者，可以直接查阅表 3.1 的内容。该表归纳了剑桥系列英语考试中不同写作测试任务所需涉及的认知活动和要求。

第 4 章　环境效度　第 4 章主要从环境效度来分析写作测试。作者认为，“环境效度与考生能胜任写作任务所需达到的语言和内容要求有关，也与描述写作行为的任务设计等特征有关”（见本书第 73 页）。因此，写作测试的环境效度验证可以从三个方面展开，即语言要求（linguistic demands）、任务设计（setting: task）和施考环境（setting: administration）。读者可通过图 4.1 参阅具体的分类。由于本章涉及的小点较多，作者采用类似问题清单的模式列举了不同环境效度的要求，并逐一比对，呈现了剑桥系列英语考试在环境效度方面的研究和举措。

语言要求包含写作任务的输入和输出两个管道。具体而言，语言要求涉及词汇资源（lexical resources）、结构资源（structural resources）、话语模式（discourse mode）、功能资源（functional resources）以及内容知识（content knowledge）。作者在解释了这些语言要求的内涵后，通过具体的举措说明了剑桥系列英语考试在这一方面的举证。比如，就词汇资源而言，剑桥大学考试委员会建立了各个考试级别的语料库，

命题人员借助语料库来命制写作试题，测试对应级别考生的词汇知识。再如，通过建立学习者语料库来纵向比较考生词汇丰富度的指标。又如，通过建立相对稳定的话题分类，使得不同级别的考试在话题难度上体现系统性的差异和衔接。

任务设计涵盖的内容较多，包括作答方式（response format）、考试目的（purpose）、评分标准知识（knowledge of criteria）、权重（weighting）、文本长度（text length）、时间限制（time constraints）和作者—读者关系（writer-reader relationship）等。其中有诸多方面值得读者参阅。比如，剑桥系列英语考试的写作测试均参照了考试开发者所研制的写作共同标准（Common Scale for Writing），将系列考试的写作能力清楚地描述给考生和其他相关方。另一亮点则是作者—读者关系。由于不同级别的考试对写作能力要求不同，本章中作者对这些考试如何界定不同的作者—读者关系（如学术英语类考试和通用英语类考试）、构建考生在作答时的读者意识均给出了值得借鉴的做法。

最后是施考环境，这与写作测试的一些考务内容相关，涉及考场要求（physical conditions）、施考统一性（uniformity of administration）和考试安全（security）等。

第 5 章　评分效度　第 5 章主要探讨评分效度。作者指出，评分效度长久以来是写作测试研究的核心问题之一。评分效度关乎写作测试的各个环节，包括评分标准（criteria/rating scale）、评分员特征（rater characteristics）、评分过程（rating process）、评分条件（rating conditions）、评分员培训（rater training）、考后校正（post-exam adjustment）、分数报告（grading and awarding）等。

评分标准是评分的基础和依据。作者在回顾了有关评分标准的分类后，通过两个案例研究，说明了剑桥系列英语考试是如何研发写作共同标准，如何修改雅思写作评分标准的。评分员特征则涉及评分

员生理、心理和经验背景等因素。这些因素对评分都有可能带来误差，继而影响评分效度。评分过程也是影响评分的重要因素，因为过程的不同会使打分产生偏颇。比如，作者指出经验丰富的评分员和新手评分员阅读写作文本的方式不同，可分为原则性二读法（principled two-scan approach）、务实性二读法（pragmatic two-scan approach）、全篇通读法（read through approach）和暂行打分法（provisional mark approach）四种[2]。评分条件包括评分时的物理环境、考生作答字迹等。

作者在本章中用了较大篇幅介绍了评分员培训的一些细节，特别是在如何保证评分员间信度（inter-rater reliability）和评分员内信度（intra-rater reliability）方面的举措。读者也可在附录 D 中找到写作评分员培训的具体方案和流程。

作者还讨论了考后校正和分数报告问题。通过统计手段，我们可以监控评分员中的某些异常打分，并对一些打分失误的情况加以修正。比如，作者提出利用多层面 Rasch 模型的方式可以发现评分员打分的松紧度和一致性，进而对分数进行校正。又如，也可以通过概化理论，对评分员的培训和选拔作出最为优化的选择。

本章最后还专门讨论了纸笔写作考试和计算机写作考试的异同。多项实证研究指出在评分结果上两种考试模式并未产生显著的差异。此外，作者还介绍了计算机辅助评阅的阅卷方式以及自动评分的阅卷方式，这些都对协调大规模写作评分、保证评分信度和效度产生了积极的影响。

第 6 章　后效效度　在第 6 章，作者主要聚焦写作测试的后效效度。在梳理有关反拨效应的主要文献后，作者指出可从三个方面来探索写作考试的后效效度，即对个体在课堂或职场中的反拨效应（washback on individuals in classroom/workplace）、对机构和社会的影响

2　限于篇幅，本导读对这些方式不逐一解释，读者可参阅本书第 198-199 页。

（impact on institutions and society）以及杜绝测试的偏颇（avoidance of test bias）。

本章主要报告了剑桥系列英语考试的一些实证研究，试图回答有关后效效度的问题。比如，考试对教材编写和开发有何影响？教材在多大程度上反映出考试的痕迹？再如，（突击）备考对考试成绩的影响如何？考试在社会和机构中的认可度如何？又如，考试对不同性别、不同民族、不同家庭背景的考生是否会产生偏见？这些研究都为有关考试中写作测试的后效效度进行了举证。

第 7 章　对标效度　第 7 章主要探讨对标效度。作者首先引用了 Weir（2005）的定义，即对标效度是一种定量性质的考后验证效度，主要涉及考分与外部标准的相关程度。一般而言，对标效度可分为共时效度和预测效度，不仅需要慎重选择所对标的外部标准，其本身也应具备较高质量。作者主要从三个方面阐述了写作测试的对标效度，即考试间的可比性（cross-test comparability）、同一考试不同版本的比较（comparison with different versions of the same test）以及与外部标准的比较（comparison with external standards）。

在研究考试间的可比性问题上，鉴于各项考试之间在性质等方面的不同，选择一项合适的外部标准来作为对标对象，这本身就极为复杂。作者汇报了剑桥大学考试委员会所开展的一系列研究工作，并将各考试的得分与 CEFR 对接，得出图 7.2 和图 7.3 的结果。而就写作测试来说，通过研发写作共同标准，不同考试的写作得分可以通过统一标尺进行衡量（读者可参阅图 7.4）。作者也指出，为了保证同一考试在不同版本上的平行程度，有的研究人员通过专家评判表（checklist）以及评分员观察（rater observation）等方式对一考多卷的平行性问题加以举证。最后，本章讨论了与不同外部标准做比较的结果以及比较过程中需要慎重处理的方面。

第 8 章　总结　作者在这一章中再次回归到 Weir（2005）的社会—认知模型，并对此进行评述。在此基础上，本章回顾了以上讨论的六个效度组成，逐一扼要地回顾了这些效度组成和有关研究，并对一些今后可以继续探索的效度研究提出了建议。

三、阅读建议

为帮助读者更好地理解本书内容，本导读给出以下三个方面的建议。

第一，全书的研究深深扎根于 Weir（2005）所提出的社会—认知模型，并对这一研究框架有了进一步的发展和丰富。读者在阅读本书之前不妨先简要地翻阅有关社会—认知模型的主要内容和基本构成，并对其中的术语有所了解。在此基础上阅读本书会收到事半功倍的效果。

第二，全书介绍了剑桥系列英语考试的有关情况，并详细说明了其中的细节。因此在阅读本书前，读者可先浏览本书的附录 A，即各项考试的样题。读者亦可在浏览的基础上先自行提炼系列考试的层级差异。这样，在阅读本书的大量实例时，读者对例子的熟悉度会较高。

第三，全书介绍了剑桥大学考试委员会很多值得借鉴的做法。从推广性的角度而言，第 5 章中有关评分效度的内容以及最后一章中的一些建议特别值得一读，建议读者重点关注。

潘鸣威

2019 年 12 月

参考文献

Chapelle, C., Enright, M. and Jamieson, J. (2008). *Building a Validity Argument for the Test of English as a Foreign Language.* London: Routledge.

Field, J. (2004) *Psycholinguistics: The Key Concepts.* London: Routledge.

Field, J. (2005) Second language writing: A language problem or a writing problem?, paper presented at IATEFL Research SIG 'Writing revisited' Conference, Cambridge, 25–27 February 2005.

Hamp-Lyons, L. (2011). Book review, *Examining Writing: Research and Practice in Assessing Second Language Writing. Assessing Writing*, 17(1), 71–74.

Hayes, J. R. and Flower, L. S. (1980) Identifying the organisation of writing process. In L. W. Gregg and E. R. Steinberg (Eds) *Cognitive Process in Writing* (pp. 3–30). Hillsdale, NJ: Lawrence Erlbaum Associates.

O'Sullivan, B. (2000) Towards a Model of Performance in Oral Language Testing, unpublished PhD dissertation, University of Reading.

Weigle, S. R. (2002). *Assessing Writing.* Cambridge: Cambridge University Press.

Weigle, S. R. (2010). Book review, *Examining Writing: Research and Practice in Assessing Second Language Writing. Language Testing*, 27(1), 141–144.

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

杨惠中、Weir, C. J. (1998). 大学英语四、六级考试效度研究 . 上海：上海外语教育出版社 .

邹申 . (1998). TEM 考试效度研究 . 上海：上海外语教育出版社 .