社会—认知视域下剑桥英语体系考试 FCE 构念界定、任务设计与评分方法

常　伟

北京秋思国际教育咨询有限公司 北京 100000

【摘　要】：剑桥第一英语证书考试（FCE）作为 CEFR B2 级别的权威认证，被广泛视为独立英语运用能力的证明。然而，在中国应试教育语境下出现了一个值得关注的现象：有人认为通过了 FCE 考试就等于中学英语高枕无忧了。本文借鉴社会—认知效验框架（Weir, 2005）和“构念—任务—评分”三维分析模型[4]，对 FCE 的构念界定、任务设计与评分方法进行系统评述。本文揭示了 FCE 的技术优势何以在产业化备考中被重构，并指出 FCE 证书的价值并非固有属性，而是教育生态与使用者测评素养共同建构的结果。研究为考试使用者、教育者及我国考试改革提供了启示。

【关键词】：FCE；构念界定；任务设计；评分方法；社会—认知框架；反拨效应

DOI:10.12417/2705-1358.26.08.020

1 引言

近年来，FCE 考生群体呈现明显低龄化趋势。受自媒体等社交平台宣传影响，“英语学习要趁早”“小学毕业前通过 FCE 可中学无忧”“搞定英语”等观点渐趋流行，其背后功利性学习动机亦获得广泛认同。

然而，笔者在调研中发现：部分小学阶段即通过该考试的学习者，初中后英语考试成绩并不理想。一方面，作为 CEFR B2 级别的权威证书，理应表征学习者具备独立运用英语进行交际的能力；另一方面，校内考试成绩不理想却对这一能力表征构成质疑。证书是否等于中学英语学习的“高枕无忧”？

带着上述困惑，本文对 FCE 的构念界定、任务设计及评分方法展开系统评述，并结合我国中高考英语考试的比较分析，探究 FCE 证书与中学英语学业表现之间的复杂关系。

2 FCE 构念界定

2.1 简介

FCE 是剑桥通用英语五级证书考试的第三级。该考试自 1939 年问世以来，其发展历程映射了语言测试理论从结构主义向交际法的范式转移[8]。FCE 定位于测量学习者在学习、工作及旅行情境中的独立语言运用能力。

2.2 构念

构念界定指一项考试对所测量的能力或内容进行界定，即回答该考试“考什么”[4]。FCE 分别测试阅读、写作、听力和口语，此外还包含对语言结构理解的考查。阅读与语用共 7 个题型、52 个题目，考查考生理解各类出版物文本的能力，用时 75 分钟；写作部分共 2 个题型，考生需在 80 分钟内完成一篇定题作文和一选择性题目作文，字数均为 140-190 词；听力部分共 4 个题型、30 个题目，考查考生听懂讲座、广播、演讲、谈话等口语材料的能力，用时约 40 分钟；口语部分由两名考生一组进行，用时 14 分钟，考查考生与考官交流、与组内其他考生互动以及单独发言时的英语表达能力。这四个部分共同评定考生在 B2 级别的整体语言交际能力（FCE 官方备考指南，2025）。

剑桥通用五级考试的理论根基可追溯至 Hymes（1972）的交际能力学说，经 Canale&Swain（1980）系统化发展，由 Bachman（1990）整合为交际语言能力模型（CLA）。辜向东、孟磊（2015）指出，剑桥系列考试始终以 Weir（2005）的社会-认知框架为构念评测依据，强调“测试任务的语境、评分标准、分数解释”三者的统一。杨吕娜、武尊民（2014）进一步总结，其优势在于将考试的技术质量（认知、情境、评分）与其社会后果（后效效度）纳入统一的效度论证体系。以 FCE 口语为例，其设计充分体现了上述构念理念。第一部分为简短的社交交流，帮助考生自然放松进入状态；第二部分要求考生进行 1 分钟不间断图片描述，考查连贯表达观点的能力；第三部分为双人或三人讨论，考查开启、转换、结束话题以及沟通协商、达成决策的能力；第四部分围绕第三部分话题展开进一步讨论，考查深入交流与任务推进能力。FCE 口语任务还原了真实交际中的不可预测性，以技能综合、认知能力、真实性为维度，强调语言能力并非孤立的知识体系，而是在具体情境中运用语言知识、策略能力、话题知识完成交际任务的能力。

3 考试任务设计

罗凯洲和韩宝成（2018）把考试设计（怎么考）比成尺子测量身高，同时考试也需要具体的任务来测量考生的能力（或表现）。FCE 作为剑桥通用英语五级证书考试的核心级别，其任务设计理念深受交际语言测试理论的影响，强调在真实语境

中综合运用语言能力（Bachman&Palmer,2010）。

3.1 任务设计的理论立场：融合视角下的综合任务

语言测试领域长期存在两种任务设计取向：一是以选择题为代表的选择式作答任务，评分成本低、一致性高，但与现实语言运用相去甚远，使用不当易给教与学带来负面反拨[5][4]；二是以开放式问答为代表的建构式作答任务，对语言运用水平的测量更为直接，但评分成本高、一致性低。FCE 的任务设计并未陷入“非此即彼”的二元困境，而是借鉴融合视角，在真实性与可操作性之间寻求平衡[4]。这一平衡的集中体现是综合任务的大量使用。

3.2 任务设计的核心原则：真实性、综合性、典型性

真实性原则。[6]曾批评国内大规模考试的“便利性偏差”——“倾向于选择方便测试的内容而不是真正重要的内容”。反观 FCE 口语测试，采用双人互动、角色扮演、协作决策等任务形式，直接测量真实交际情境中的口头互动能力，体现了真实沟通情境、沟通结果导向、聚焦意义协商三大原则（Ellis，2019，转引自[4]。综合性原则。FCE 在多板块任务中实现了技能整合：写作 Part 1 的读写、听力 Part 2 的听写、口语 Part 3 的听说联动，共同构成技能综合的任务矩阵。典型性原则。FCE 的任务类型是对目标语言使用情境中高频、高价值交际任务的取样，这一取样策略正是内容效度的核心保障[6]。4.评分方法

“如何评”涉及评分方式的选择、评分标准的确定、评分过程的质量监控以及分数体系的构建[4]。

4 评分方法

4.1 评分方式：分项评分与整体判断相结合

FCE 写作与口语任务均采用建构型作答形式，由经过专业培训的评分员进行人工评分。为兼顾评分的信度与构念覆盖度，FCE 采用分项评分与整体判断相结合的模式。评分员首先从若干分项维度对考生的表现进行评价，再基于整体印象形成综合判断。正如[1]所言，剑桥系列考试历来注重通过三项措施来保障评分一致性：增加写作任务、引入多人评分、采用分析式评分。FCE 写作评分从“内容”“沟通水平”“条理性”“语言”四个维度分别赋分（剑桥官方手册，2015）。

4.2 评分标准：等级描述与分项细则

评分标准是评分行为的依据。FCE 的评分标准采用五档量表（0-5 档），每档附有详细的描述语。例如，写作的“内容”维度在 5 分档要求“所有内容切题。目标读者可以充分获悉有关信息”；“语言”维度在 5 分档要求“恰当使用一系列词语，包括较高级的词语。能够正确、灵活地运用多种简单的和复杂的语法形式。可能存在个别错误，但不影响意思传达”。这些描述语直接源于 CEFR B2 级别的“能做”描述，体现了构念与评分的统一。[2]在评介剑桥听力测试研究时曾高度评价剑桥的分数报告体系，指出“考生得到的信息包括标准分、等级、考生在每一部分的表现以及考生的强项与弱项”，FCE 其分项评分为考生提供了诊断性反馈，使其能够识别自身在各项技能上的优势与不足[7]。

4.3 评分过程：双评机制与质量监控

为确保评分的公平性与一致性，FCE 所有写作和口语任务均采用双评机制。两位评分员独立评分，若分差超出预设阈值，则自动提交给第三位资深评分员（评分组长）仲裁。评分前，所有评分员需接受严格培训，并使用标准卷（benchmarking）进行校准；评分过程中，系统会随机插入已评定的标准卷以监控评分员的一致性[1]。[3]提醒，过度追求评分一致性可能带来副作用——评分员可能因此放弃独立、主观的专业判断，转而依赖易于识别的表层特征进行简单评判。评分培训因此特别强调对“互动交流”“话语组织”等深层构念要素的关注，力求在信度与构念效度之间取得平衡。

4.4 分数体系：量表分、等级与诊断报告

FCE 的成绩报道采用标准参照方式，分数与 CEFR 等级直接挂钩。考生获得的不是原始分，而是经过等值处理的量表分（scale score），范围在 140-190 之间。总分 160 分以上对应 B2 等级（剑桥官方手册，2015）。除总分外，考生还收到各技能的分项得分，以及每个分项维度的表现描述。[2]指出，“这种诊断性与解释性并重的分数报告模式，是国内大规模考试亟待借鉴之处”。然而[7]，也提醒，分项评分本身并不自动产生教学价值，分数的解释与使用才是关键。

5 结语

FCE 构念以 CEFR B2 级别为基准，任务设计遵循真实性、综合性、典型性原则，评分体系采用分项评分与整体判断相结合的模式，分数报告兼具诊断性与解释性[2][1]。这些设计使其在测量学习者综合语言运用能力方面具有显著的科学合理性，也为我国中高考改革提供了可贵参照。

然而，将 FCE 与国内中高考相比，FCE 在技能覆盖面、任务真实性和评分诊断性上均领先一步；但正因其任务综合、技能全面，也更容易被应试产业“拆解”——培训机构将综合性任务还原为分项技巧，使考试所倡导的“真实沟通能力”在备考实践中悄然异化。桂诗春教授曾深刻指出：“不少大规模考试早期都能起到一些好的反拨作用。可时间一长，对付考试的办法多了，反拨作用也就变了。”[5]这一异化在特定群体身上尤为显著：中学后英语成绩并不理想。从内容效度视角看，FCE 备考若以真题训练为核心，学习者的语言输入便被窄化为

考试抽样内容的反复强化，而非 B2 所要求的广泛真实的语言接触[6]。从反拨效应视角看，考试的影响亦受教师素质、学校理念、家庭教育资本等非考试因素调节

因此，对 FCE 的评价不应止步于技术层面的赞誉或质疑。它提醒我们：一张考试证书的社会价值，始终是特定评价制度与文化情境的产物。正如[4]所言：“测评最终能否促学，并不在于是何‘类别’，而在使用的人能否对测评本质有精准理解、能否对测评决策有合理运用。”

参考文献：

[1] 辜向东,孟磊.(2015).剑桥英语百年测了什么？《构念评测:剑桥英语测试百年史》述评.外语测试与教学,(4),59-64.

[2] 何莲珍.(2019).《第二语言听力测试研究与实践》评介.外语教育研究前沿,2(3),80-84.

[3] 金艳.(2019).社会—认知视角下的口语测试效度研究——《第二语言口语测评研究与实践》评介.外语教育研究前沿,2(2),81-84.

[4] 罗凯洲,韩宝成.(2018).国才考试的构念界定、任务设计与评分方法.中国外语教育(季刊),11(1),40-46.

[5] 亓鲁霞.(2004).NMET 的反拨作用.外语教学与研究,36(5),357-363.

[6] 宋得龙,陈黎萍.(2019).基于效度理论的中考英语试卷命制策略与方法.中小学英语教学研究,(360),48-58.

[7] 徐加永,罗凯洲,王佳雨.(2024).高考英语邮件写作分项评分质量研究.西安外国语大学学报,32(4),80-85.

[8] 杨吕娜,武尊民.(2014).剑桥考试中心语言测试研究系列简述——纪念剑桥英语考试 100 周年.外语测试与教学,(4),47-61.