Frontiers in Social Sciences and Humanities 社科与人文前沿, 2024, 2(1), 1-6
Published Online June 2024.
http://www.ufascience.com/journal.aspx?ProID=5

传统监考模式与在线监考模式口语考试的对比研究

张杰

大连理工大学，大连 116024，中国

摘要：本文探讨了传统的中心模式 (TM) 和在线监考模式 (OLP) 下面对面进行的高风险英语口语测试成绩的可比性。数据包括大量通过 TM 或 OLP 参加四个 CEFR（“欧洲语言共同参考框架”）级别（B1 至 C2）的英语口语测试的考生样本。使用描述性统计、效应大小差异和等效性检验对数据进行了分析。虽然在 C2 级别，两种模式获得的分数存在一定差异，但这些差异并不具有统计学意义。本文的结论是，无论口语测试采用在线监考模式还是传统的面对面模式进行，考生获得的分数都相似。研究证实，考试方式不会显著影响考生的分数。

关键词：考试成绩可比性；英语语言；口语测试；CEFR；在线监考

The Delivery of Speaking Tests in Traditional or Online Proctored Mode: A Comparability Study

Jie Zhang

Dalian University of Technology, Dalian 116024, China

Abstract: This paper investigates the comparability of test scores recorded for high-stakes English language Speaking Tests administered face-to-face in either a traditional centre-based mode (TM) or in an online proctored mode (OLP). The data comprise a large sample of test takers taking English language Speaking Tests at four CEFR (the ‘Common European Framework of Reference for Languages’) levels — B1 to C2 — via TM or OLP. The data were analysed using descriptive statistics, effect size differences and equivalence tests. While a degree of difference in scores obtained between modes was apparent at C2 level, the differences were not found to be statistically significant. The paper concludes that whether Speaking Tests are delivered in online proctored mode or in traditional face-to-face mode, test takers receive similar scores. The study confirms that mode of test delivery does not significantly affect test taker scores.

Keywords: Test score comparability; English language; Speaking tests; CEFR; Online proctoring

自 2010 年代末以来，以及最近由于新冠疫情的影响，许多考试已从面授转为在线授课。本研究旨在确定授课模式对考试成绩的影响程度，进而影响口语考试成绩。本文以欧洲共同参考框架（CEFR）B2 至 C2 级别的英语口语测试为研究对象，考察传统面授模式（TM）与在线监考模式（OLP）考试中考生所获分数的可比性。

Copyright © 2024 by author(s) and Upubscience Publisher.
This work is licensed under the Creative Commons Attribution international License (CC By 4.0)
http://creativecommons.org/licenses/by/4.0/
OPEN ACCESS

文章引用: 张杰. 传统监考模式与在线监考模式口语考试的对比研究[J]. 社科与人文前沿, 2024, 2(1): 1-6.
DOI:

本文首先回顾了日益普遍的在线学习和教学方式。随后，本文回顾了不太常见的在线考试方式。最后，本文简要探讨了口语评估以及开展交际口语测试的挑战。最后，本文探讨了比较这两种授课模式的研究。

在介绍背景之后，我们将展示通过 TM 和 OLP 参加 CEFR B1 至 C2 级别英语口语测试的大量考生样本数据，并分析其统计差异。

一、背景

本节介绍在线学习与教学的背景，尤其是在新冠疫情 (Covid-19) 疫情期间。随后，本文探讨了在线评估中存在的问题——在线学习语言 (OLP) 模式下考试的利弊。此外，本文还简要探讨了口语评估，以及口语沟通能力评估中面临的特殊挑战。最后，本文探讨了日益棘手的在线口语评估问题。

面对新冠疫情，教师在课堂上进行教学的常见做法发生了巨大而迅速的变化 [1]。随着科技的发展，在线学习的接受度在过去两年中呈指数级增长[2]，人们开始重新思考“传统”的授课模式 [3]。例如，Todd 概述了新冠疫情如何有力地推动了在线教学的普及。

尽管人们对在线授课内容的观念发生了变化，但考试仍然被视为一种在更传统的面对面授课环境中进行的活动[4]。评估领域已开始采用技术，但其应用程度远不及在线教学[5]。

评估——尤其是某些在线测试普遍存在的公立学校系统之外的高风险评估——通常被视为在实体考试中心以纸笔模式在考官/监考人员面前进行的活动。虽然在线学习技术使得学习和教学相对高效，但在线评估模式也存在诸多优势、问题和挑战：例如，作弊行为减少、网络连接问题等[6]。

Khan & Jawaid 报告了新冠疫情期间巴基斯坦的在线评估情况，探讨了在获取和交付方面，学习、教学和评估尤其需要得到同等重视，并强调在线评估需要转变观念，让经济发展中国家的管理人员和考生都不再惧怕新技术。

García-Peñalvo 等结合西班牙大学应对新冠疫情的情况，提出了一些有关在线评估的建议。除了加强持续性评估外，他们还建议应使用支持面对面教学的技术（例如电话会议）进行评估，以培养教师和学生对“在线评估的新环境”的准备度和信心。他们强调，任何评分方案都必须在进行任何评估之前告知学生。García-Peñalvo 等人建议，当涉及“涉及大量学生的复杂科目”时，应针对相关科目或学生群体开发专门设计的在线评估方法。

正如 Weiner 和 Henderson 所指出的，以在线学习模式（OLP）进行考试对考生和考试机构来说既有好处也有坏处。积极的一面是，考生可以在自己舒适（且安全）的家中参加在线监考考试，这在疫情期间行动受限或残障考生难以进入远程考试中心（即使在正常情况下，更不用说在疫情期间）的情况下是一个重要因素。此外，考试的发送和成绩的发布速度也可能体现了在线考试模式的优势。

与在线评估相比，在线教学在实践和预期方面的历史更为悠久。在线教学的研究已有十余年，强调协作原则，例如讨论、同伴支持、因材施教、自主学习、鼓励学生设定目标以及规划、监控和控制自身的认知[7]。相比之下，在线评估的记录较短。在线评估（尤其是高风险评估）的预期仍然较为传统，并且直到最近，通常都是由一位考生独立完成的。此外，在考试交付方面，传统的可比性（以及由此产生的信度）观点通常要求所有考生同时获得相同的评估结果。然而，在网络世界中，传统的大规模评估方法难以实施，这样的要求可能会引发安全性、诚实性和公平性方面的问题。

关于在线考试 (OLP)，围绕安全性、在线考试的“脆弱性”以及学术不诚实等问题已经展开了广泛的讨论[8]。这些问题至关重要，尤其是在考试在考生家中等远程地点进行的情况下。

尽管如此，Foster 和 Layman 描述了如何设置安全级别，使在线考试监考切实可行。事实上，已有研究报

告称，与传统的面对面考试相比，在线考试监考技术的应用甚至可能使考试安全性更加有效。

技术因素也可能需要加以考虑。在对 OLP 考试的评估中，Giller 等人提出了一些建议，建议将在线考试监考技术应用于在线考试，并评估其有效性。

尽管存在这些担忧，OLP 在未来仍然是一种潜在的重要交付方式。本研究探讨了 OLP 口语评估与传统方法的可比性及其互换性。

本文将简要概述口语技能评估和远程评估的关键问题。

长期以来，口语一直被认为是四项宏观技能中最难评估的。大约 40 年前，Madsen 概述了口语难以评估的一些原因。除了背景构建问题（例如定义口语技能的实际性质和制定在交际时代正确评估口语的标准）之外，还必须处理能力、语气、推理等因素，以及一些考生甚至不愿开口说话。

Luoma 重申了口语是最难可靠评估的语言技能。尤其是在面对面交流中，由人工评估员进行口语评估时，评估结果会受到诸多因素的影响，例如口语的特征、考生的语言水平、性别、互动的性质、互动中涉及的任务和主题，以及考生展示其能力的机会。

Sujana 在讨论口语能力测试的复杂性时，也表达了上述许多观点，并指出许多教师几乎避免评估口语能力。

如上所述，口语评估涉及各种“复杂性”。为了克服这些复杂性，许多教育工作者和研究人员建议将口语评估转移到在线模式，他们认为这比面对面模式更具优势。例如，Fall 等人 描述了一种模拟口语能力面试 (SOPI)，它使得根据 ACTFL 口语能力量表对考生口语能力进行大规模评估变得相对容易管理和评分。然而，该过程完全由机器介导。

在新冠疫情的背景下，各种形式的评估都转移到了各种在线模式，并取得了不同程度的成功 。正如预期的那样——根据上文关于口语评估复杂性的讨论——评估学生的口语能力确实是许多教育工作者面临的最大挑战。Forrester 详细阐述了新冠疫情期间在线口语评估的挑战。这些问题适用于所有形式的口语能力评估，而不仅仅是正式考试。

二、OLP/TM 考试结果的可比性

目前已有大量关于在线监考和无监考评估的研究，但很少有研究直接比较 OLP 的高风险考试与传统的中心面对面模式的考试。下一节将简要介绍这两个相关但不同领域的研究。

许多关于不同监考模式的研究都发生在高等教育领域。在高等教育之外和组织心理学领域，Tippins 探讨了新技术如何导致“对良好考试实践假设的改变”以及“需要面对技术进步带来的新问题”。她还举例说明了技术如何以现实的方式应用于评估。总体而言，研究表明，参加无监考考试（无论远程或其他方式）的学生比参加远程监考考试的学生成绩更高（这或许并不令人意外）。

然而，也有研究报告称，无论有无监考，学生的考试成绩并无显著差异。

尽管在 2020-2022 年新冠疫情之后，在线进行的高风险评估有所增加，但正如 Weiner & Henderson 所观察到的，关于远程监考考试与在考试中心面对面监考考试中获得的高风险考试成绩的可比性的研究却很少。下文概述了该领域有限的研究成果。

Weiner 和 Hurtz 在美国执照考试的背景下考察了考生的表现，探索了在特制的配备计算机的“自助服务终端”中参加考试的考生与在有人工监考的实体考试中心参加相同考试的考生的表现在多大程度上相同。在两种监考模式下，考生的表现均未发现显著差异。在新冠疫情导致考试长期关闭后，Hurtz 和 Wiener 扩大了上述研究的范围。他们的研究报告称，监考模式并未对考试成绩产生影响。

Wuthisatian 研究了使用远程在线监考的考生与在传统考试中心参加高风险经济学考试的考生之间的表现差

异。结果表明，考生在两种监考方式中的表现不同：在中心参加考试的考生获得的分数明显高于接受在线监考的考生。Cherry 等人研究了美国的专业执照考试，比较了远程在线监考和在考试中心进行的考试结果。虽然两种模式的结果存在统计学上的显著差异，但并未发现任何一种模式更有利的模式。

Morin 等人调查了加拿大一项重要的全国医师执照考试，该考试通过远程在线监考或考试中心进行。尽管一些考生报告了不同的考试体验，但 Morin 等人报告称，两种监考模式下的考试成绩——尽管考试题型不同——大致相当。

Muckle 等人的研究探讨了新冠疫情后通过两种监考模式进行的北美药剂师执照考试的分数。Muckle 等人报告称，考生在现场参加的考试分数更高。虽然他们将部分结果差异归因于样本构成，但显然需要进一步研究。针对考生对 LanguageCert OLP 测试方式的反应进行的研究迄今为止总体上是积极的——与 Muckle 等人在其研究中报告的结果大致相同

三、研究

本研究的数据来自 LanguageCert 于 2019 年至 2021 年间实施的国际英语学习者 (IESOL) 口语测试系列，该系列测试中的每项测试均与欧洲语言参考框架 (CEFR) 级别保持一致。LanguageCert 口语资格认证包含一项全面的英语口语测试，其考试任务旨在评估英语在现实生活中的运用能力。该资格认证适用于全球非英语母语人士；在英国或海外参加英语课程的青少年或成年人；将英语作为学校或大学课程一部分的学生；以及申请来英国工作的人士。

所有口语测试均包含四项任务，随着考生的深入，任务难度逐渐增加，B1 考试时长为 12 分钟，C2 考试时长为 17 分钟。测试共有四个评分量表，每个量表包含四个评分等级。口语测试由现场对话者进行（无论是面对面还是远程监考），所有考试内容均会被录制，以便日后评分和申诉。所有口语测试均根据四个评分标准进行评分。满分为 50 分，评分等级如下：低于 50% 为不及格；50%-74% 为及格；75% 及以上为高分及格。详情请参阅 https://www.languagecert.org/en/language-exams/english/languagecert-international-esol。

所有考试均由 LanguageCert 的封闭式阅卷员进行评估，阅卷员会定期接受培训，以确保评估的一致性和客观性，并以欧洲共同参考框架 (CEFR) 为基准。每个级别的考试都提供多种不同的试卷，并且新的试卷会不断添加到试卷库中。

为了增强安全性，不仅不同的测试表格是随机使用的，而且构成测试表格的四种任务类型也是随机的。

LanguageCert 在全球范围内运营 OLP，其测试遍布全球 70 多个国家/地区。因此，OLP 进行的所有环节——登录、安全检查、连接和语音质量检查等——均通过英语进行。考虑到低水平测试者可能存在的英语语言限制，OLP 进行的 IESOL 测试主要针对 B1 及以上级别。因此，以下口语数据集仅适用于 CEFR B1 至 C2 级别。

在测试信度方面，由于口语测试分数是通过四个评分量表获得的，因此无法通过基于项目或评分者的估计方法来评估信度。但是，可以通过单维因子分析来估计信度，该分析通过四项宏观技能（即阅读、听力、写作和口语）的原始总分以及授予的 CEFR 等级来计算麦当劳欧米茄 (McDonald's Omega)。可以看出，所有 CEFR 级别的口语测试负荷和成绩都在 0.90 及以上，这表明口语测试具有较高的可靠性。

等价独立样本 t 检验允许用户检验零假设，即两个独立组的总体均值落在用户定义的区间（即等价区域）内。使用双单侧检验（TOST）的程序可以通过指定的上限和下限来观察显著性，这与报告单个 t 值的标准 t 检验不同。正如 Lakens 所述：

采用等价检验可以避免将不显著的 p 值误解为没有效应，并促使研究人员明确他们认为哪些效应是有价值的。

上限和下限表示被测样本的两个总体 t 值的变异程度。如果等效性检验的 t 值在估计范围内，则可以认为两个群体等效。

本研究的总体假设是，两种测试模式（OLP 和 TM）之间获得的平均分数不会有显著差异。具体而言，本研究基于以下两个假设：

(1) 最坏情况下，两种模式之间也只会观察到很小的效应量差异。

(2) 在等效性检验中，对于任何给定的 CEFR 等级，在指定的上限和下限范围内，显著性都不会显现。

上限和下限设定为原始分数的+/- 0.05（即 95%区间）。这些界限可以理解为代表 95%的置信区间；然而，由于 TOST 由两个单侧检验组成，因此更准确地说，指的是置信区间的上限和下限。如前所述，等效性的关键判定在于估计的 t 值是否介于上限和下限之间。t 值（上限、T 检验和下限）的 p 值表示 T 检验值在超出指定范围时具有显著性。

在四个级别中，均未观察到下限和上限的显著性。这表明，尽管并非完全匹配，但对于本研究中所有 CEFR 级别，两种口语测试模式可以视为大致等效。话虽如此，C2 级别测试似乎存在问题，显然需要进行更多研究。

四、讨论与结论

本研究探讨了 LanguageCert IESOL 英语口语测试（CEFR B1 至 C2 级别）考生通过传统面授模式 (TM) 和在线监考模式 (OLP) 获得的分数的可比性。

本研究的关键假设是，在 OLP 和 TM 两种测试模式下获得的平均分数以及由此产生的成绩不会有显著差异。具体来说，本研究检验了两个假设。

第一个假设是，在最坏的情况下，两种模式之间也只能观察到很小的效应量差异。虽然在 B1 至 C1 级别观察到的效应量可以忽略不计，但在 C2 级别观察到的效应量为中小规模，这意味着该假设无法成立。

第二个假设是，在等效性检验中，对于任何给定的 CEFR 级别，在指定的上下限范围内，显著性都不会显现。由于在任何测试级别中均未观察到两个界限的显著性，因此确定在所考察的四个 CEFR 级别中，两种考试管理模式大致上可以视为等效，并接受了该假设。然而，在最高能力水平（CEFR C2 级别）下，考生在在线监考模式下的得分明显高于面对面监考模式。

造成这种差异的原因可能有两个。其一与 C2 考生群体的实际构成有关。C2 级别的考生往往是 30 多岁和 40 多岁的专业人士，而在较低级别的考生中，许多考生是更年轻的学童，他们更习惯于传统的面对面中心评估。因此，C2 考生也更适应广泛使用技术，这一事实可能解释了他们在在线监考环境中可能更自在的原因。第二个问题可能是不当行为。然而，在口语测试之前，我们会进行严格的安全检查，以防止诸如冒充之类的问题。如上所述，口语测试的材料是随机的，以防止可能出现的预先安排好的答案。此外，口语测试是一项实时进行的口语能力测试，从考生的角度来看，这使得作弊变得更加困难。

总而言之，从参加较低 CEFR 级别的 LanguageCert IESOL 口语测试的结果来看，无论是传统的面授模式还是在线监考模式，都会获得类似的结果。尽管如此，C2 级考生得分更高的事实确实需要在这一级别上进行进一步的研究。

本研究的一个局限性在于仅研究了一项技能——口语。口语技能通常被认为是最难管理和评估的技能，而在线授课的难度比听、读、写等较为“静态”（指不需要与对话者直接互动）的技能更甚。一项针对听、读、写等其他技能的后续分析研究正在进行中。

参考文献

[1] Hodges C, Moore S, Lockee B, et al. The difference between emergency remote teaching and online learning. Educause Review, 2020.
[2] Lim C P, Wang L. Blended learning for quality higher education: Selected case studies on implementation from Asia-Pacific. Bangkok: UNESCO Bangkok Office, 2016.
[3] Todd R W. Teachers'perceptions ofthe shift from the classroom to online teaching. International Journal of Tesol Studies, 2020, 2(2): 4-16.
[4] Coniam D, Lampropoulou L, Cheilari A. Online proctoring of high-stakes examinations: A survey of past test takers' attitudes and perceptions. English Language Teaching, 2021, 14(8): 58-72.
[5] Gardner L. Covid-19 has forced higher ed to pivot to online learning. Here are 7 takeaways so far. The Chronicle of Higher Education, 2020, 20(5).
[6] Mays T J. Teaching the teachers. In Radical Solutions for Education in a Crisis Context. Springer, Singapore, 2021: 163-176.
[7] Sarrayrih M A, Ilyas M. Challenges of online exam, performances and problems for online university exam. International Journal of Computer Science Issues, 2013, 10(1): 439.
[8] Berrada K, Ahmad H A S, Margoum S, et al. From the paper textbook to the online screen: A smart strategy to survive as an online learner. In Radical Solutions for Education in a Crisis Context. Singapore: Springer, 2021: 191-205.