

A Review of the Validity and Reliability of Chinese Oral Test in the Past Two Decades

Mongkhon Chaiarram¹; Zhiguo Wang²

School of Liberal Arts, University of Phayao, Phayao 56000 Thailand

Email: xuyong0810@gmail.com

(received 22 October 2024)(revised 13 December 2024)(accepted 14 January 2025)

Abstract

With the growing demand for teaching and assessment of Chinese as a Second Language, oral proficiency testing has received heightened attention as a key component affecting educational quality. Oral Chinese Test of international Chinese education is becoming more and more important nowadays due to the implementation of the *Chinese Proficiency Grading Standards for International Chinese Education*. This paper reviews the literature on Chinese oral proficiency testing research since the beginning of the 21st century, analyzing the dynamics and progress of validity and reliability studies within this domain. The study selects relevant literature from authoritative databases for analysis and uses a literature review method to organize and assess the scientific rigor and reliability of various oral proficiency tests. The results indicate that while Chinese oral proficiency testing has progressively refined its scoring systems and criteria based on the principles of comprehensiveness and communicativeness, there remains room for improvement in the practical operation regarding validity and reliability. Further optimization is needed through diversified research methodologies. The paper suggests that future research should delve into the scope of applicability, evaluation standards, and technical support of Chinese oral proficiency testing to enhance its overall quality.

Keywords: Chinese Oral Test; Validity Analysis; Reliability Assessment; Literature Review

¹ Chinese Lecturer

² Lecturer, Corresponding Author

近二十年汉语口语测试的效度与信度研究现状

赵丰³、王志国⁴

泰国 碧瑶府 碧瑶大学 人文学院 邮编 56000

邮箱: xuyong0810@gmail.com

收稿日期 2024. 10. 22 修回日期 2024. 12. 13 接收日期 2025. 1. 14

摘要

随着汉语作为第二语言的教学与评估需求日益增长,口语测试作为影响教育质量的关键环节备受关注。汉语口语测试随着《国际中文教育中文水平等级标准》的实施,也变得越来越重要。本文通过梳理自 21 世纪初以来汉语口语测试研究文献,分析了该领域效度与信度的研究动态及进展。本研究选取与研究相关的文献作为分析对象,并采用文献综述法整理评估各项口语测试的科学性和可靠性。研究结果显示,汉语口语测试在坚持综合性和交际性的原则基础上,逐步完善了评分体系与标准,但在具体操作过程中仍存在效度与信度上的不足,需要通过多元化研究方法进一步优化。本文指出,未来研究应针对汉语口语测试的适用范围、评价标准以及技术支持等方面进行深化,以提高口语测试的整体质量。

关键词: 汉语口语测试、效度分析、信度评估、研究现状

1. 引言

随着国际交流的不断深入以及“一带一路”倡议的推进,汉语已成为许多国家重视的第二语言。作为汉语能力评估体系中的重要组成部分,口语测试在语言教育及评估中扮演着至关重要的角色。随着《国际中文教育中文水平等级标准》在 2021 年 7 月 21 日的正式实施推广,汉语水平考试迎来了一系列重大变化。特别是从 2023 年 1 月起,汉语水平口语测试迎来了重大改革,根据文件规定,汉语水平考试 HSK 与 HSKK 合并实施。由此可见,汉语口语测试已经变得越来越重要,成为了国际中文教学众多测试环节中必不可少的一部分,对汉语学习者的学习产生了重大影响。

汉语口语测试不仅影响教学的方向和质量,也关系到测试者语言运用能力的真实反映。因此,保证口语测试的效度和信度,对于促进汉语教学水平的提升以及国际汉语推广具有重要意义。本研究通过对自 21 世纪初以来汉语口语测试的效度与信度研究文献

³ 赵丰,男,泰国碧瑶大学中文专业讲师。

⁴ 王志国,男,泰国碧瑶大学文学院讲师。

的梳理，总结了此领域的研究进展与现状，旨在揭示汉语口语测试在理论发展与实践操作中存在的问题及挑战，进而提出可行的改进策略。这些问题不仅触及如何科学、准确评估和反映口语能力这一核心问题，也关联到测试设计的科学性、合理性及测试结果的普适性和可靠性。

本文的贡献在于系统梳理了汉语口语测试效度与信度研究的历史脉络，概括了当前主要的汉语口语测试方法与技术，并明确指出了此领域研究的不足之处。通过对现有文献的综合分析，本文不仅加深了对汉语口语测试评价标准与操作规程的理解，而且对如何提高汉语口语测试的可靠性和有效性提出了具体的研究方向。本研究的结论与建议将为今后汉语口语测试的研究及实施提供参考，对进一步推动汉语教育学科的发展具有积极影响。

2. 汉语口语测试的演变历程

2.1 二十年间汉语口语测试的发展概况

汉语口语测试是汉语学习者语言学习能力最直接的外化表现，主要考察汉语学习者的“说”的能力，即语言输出的过程，同时包括命题设计、测试实施、口试评分和成绩分析等相关过程。自 21 世纪初以来，汉语口语测试作为对外汉语教学评估体系中的重要组成部分，其发展经历了从初步探索到逐步成熟的过程。Xu (2005) 将对外汉语口语成绩测试归类于对留学生每学期期中和期末进行的口语考试，一般就是由对外汉语教师自行编写试卷并组织实施最后进行统计评分和成绩评分。但相关研究不仅关注测试内容的编制和评价标准的制定，更加注重测试的效度与信度，以及评分的公正性和科学性。

在早期阶段，汉语口语测试的发展主要受到传统语言测试理论的指导，注重语言知识的考核，而对语言运用能力的测试相对薄弱。这一时期的口语测试往往采用直接或间接的方式，如看图说话、回答问题等，评分标准多侧重发音、语法和词汇等语言形式要素，对考生的交际能力和实际运用语言的能力考察不足 (Chai, 2003)。

随着交际语言教学理念的普及和认可，汉语口语测试开始强调语言的交际功能和实际运用能力。Yang (2020) 认为这一阶段的测试设计更多地模拟真实的交际情境，例如情景模拟、角色扮演等形式，通过这些更贴近真实生活的测试形式，来检测学习者的语言运用能力。评分标准也开始考虑语言的流畅性、准确性、适切性等因素，更全面地反映考生的口语交际能力。

此后，随着测试理论的不断发展和计算机技术的应用，汉语口语测试的设计和评分过程越来越科学化、规范化。计算机辅助语言测试 (CALL) 的引入，使得测试内容更新更加及时，评分更加迅速和客观。例如，采用语音识别技术进行发音评分，利用大数据

分析评价考生的口语水平等，这些技术的应用提高了测试的效率和客观性。（Huang & Wang, 2019）

尽管汉语口语测试在设计和实施方面取得了一定的进步，但在效度和信度方面仍存在诸多挑战。Xu & Liang（2020）在研究中表明，汉语口语测试中存在着题目设计的主观性、评分标准的不统一、评卷者之间的差异等问题，这些都可能影响到测试结果的可靠性和有效性。此外，由于口语测试往往需要考虑文化背景、语境变化等复杂因素，使得测试的效度评估变得更加困难。

为了提高汉语口语测试的效度与信度，学者们提出了多种改进建议。一方面，测试编制者应当基于实际语言使用的需求来设计测试内容，确保测试任务与真实交际活动相吻合，以增强测试的效度。另一方面，测试评分过程中应当采用多位评卷者，通过训练提高评卷者间的一致性，同时采用标准化的评分细则和标准，减少主观性对评分的影响，以提高测试的信度。（Chai & Wang, 2019）

与此同时，随着语料库和大数据分析技术的发展，研究者们开始利用真实的语料来分析和构建测试内容，这些基于真实语料的测试能更准确地反映考生的语言能力，从而提高测试的效度。Kane（2013）认为，通过分析真实语料，研究者能够识别出日常交际中常用的语言结构、词汇和语用模式，并根据这些数据设计更为真实和贴近实际语言使用的测试任务。同时，通过对大量测试数据的分析，可以发现评分过程中的问题并及时调整，以提高测试的信度。例如，运用机器学习算法和统计模型，研究者能够自动识别和调整评分中的不一致性，这种技术的应用不仅提高了评分过程的效率，也保障了评分的公正性和稳定性（Graesser et al., 2004）。

总体而言，汉语口语测试的发展显示出由形式向功能转变，由传统向现代化技术应用转变的趋势。未来，汉语口语测试需要在确保效度与信度的同时，不断探索更为科学、合理的测试方法，满足不断增长的对外汉语教学评估需求。

2.2 汉语口语测试模式及其特点

在近二十年的汉语口语测试研究中，汉语口语测试的主要模式主要包括直接口语测试和半直接口语测试两种，它们各自具有独特的特点和应用场景。Chai（2003）认为，口语测试是对被试口头语言表达能力的考核，属于主观性测试题。Weir（2005）根据测试的形式，把口语测试分为三种，直接口试、半直接口试和间接口试。

从目前的口语测试和形式和方法来看，直接口试和半直接口试是普遍使用最多的测试方式。国际上常用的英语考试——雅思考试（International English Language Testing System）是最经典的直接口试，评分标准和等级判定经过若干年的发展已经非常完备。而对汉语学习者口语能力进行考察的标准考试——HSKK 是典型的半直接口试，考生并不会直接面对考官，而是根据测试指导和相关测试提示，每位考生在相同的时间

内，按照统一的标准考试程序，完成相同的考试任务。直接口试则模拟了真实的语言交际环境，但存在一些问题，如可行性差、评分信度不高的缺点；半直接口试更适合大规模标准化测试，但没有体现口语测试的真实性原则。（Phakiti, 2008; Chai, 2003; Xu, 2005）

直接口语测试的问题在于实施难度较大，对考官素质要求高，同时评分的主观性较强，可能影响信度。半直接口语测试无法全面评价考生的即时口语反应和交际能力，因此在效度方面存在一定限制。在两种主要汉语口语测试模式的基础上，Zhai (2012) 还提出了多样化的汉语口语测试方法，以期提高测试的效度和信度。一种是采用多任务测试法，即在一次口语测试中设置多种不同类型的任务，如叙述、描述、论述等，旨在全面考察考生的口语能力。这种方法有助于减少单一任务类型可能带来的评分偏差，提高评分的公平性和可靠性。

另外，随着技术的发展，计算机辅助口语测试（CAT）也逐渐成为研究热点。它通过计算机技术记录和评估考生的口语表现，不仅提高了评分效率，还通过算法减少了主观评分的差异，从而增强了信度。但由于技术限制，CAT 在评价考生口语的自然性和交际策略方面仍有待完善（McNamara, 2000）。与人工评分者不同，计算机系统可能难以全面评估语言的语调、即兴性和交际互动的动态过程。此外，语音识别和分析的技术限制也可能影响评估复杂语言使用的准确性，例如讽刺或细微表达的理解（Lu, 2014）。

虽然汉语口语测试模式不断发展和创新，但在实际操作过程中，测试的效度与信度仍然存在不足。例如，直接口语测试中考官的主观判断可能影响评分的一致性，而间接口语测试可能无法全面反映考生的实际交际能力（Zhao, 2012）。此外，不同测试模式往往针对的考生群体和语言水平有所不同，这就要求在设计和应用测试时，能够准确考量目标群体的特点和需求。

在未来的研究中，为了提高汉语口语测试的效度与信度，研究者需要从多方面进行深入探索。首先，加强考试评分标准的科学化和规范化，培训专业的考官队伍，以提高评分的客观性和准确性（Wang, 2002）。其次，结合语言测试理论和二语习得研究，对测试任务设计进行优化，使之更能反映实际的语言使用情况。再次，利用现代信息技术，探索更多元的测试手段，如虚拟现实技术在口语测试中的应用，以增强测试的互动性和趣味性。最后，增加对口语测试效度和信度的跟踪研究，通过长期追踪考生的语言学习和使用情况，验证口语测试的预测效度。

总体而言，汉语口语测试作为评价汉语作为第二语言学习者语言能力的重要手段，其科学性和可靠性对于教育评估具有重要意义。通过不断地研究和改进，可以使汉语口语测试更加完善，更好地服务于汉语教学和学习者的实际需求。

3. 效度研究梳理

3.1 汉语口语测试效度分析研究进展

语言测试的评价标准一般分为信度、效度和实用性三个方面。效度和信度是语言测试中两个非常重要的概念，是衡量语言测试质量的重要依据，是本文将要重点梳理的内容。随着语言交际能力理论不断发展，语言测试内容越来越倾向于对测试者语言实际应用能力的测量，测量的方式也逐渐倾向于测验任务的真实性，如对口头表达和交际能力的测试等。因此，口语测试的效度研究成为近年来重要的热点之一。

效度，是衡量测试的一个重要标准，简而言之就是语言测试的有效性，是指该测试是否测试出出卷人想要检测的内容、达到预期的测试目的。它是测试最基本的出发点。在研究的早期阶段，学者主要围绕汉语口语测试的命题效度进行研究。口语测试的命题效度指的是测验的题目对有关内容范围取样的适当性，也就是说测验所选的测试题目是否符合所要测量的口语能力，其代表性是否适当等。Zhao (1999) 认为设计具有高效度的口语试题需要遵循以下原则：一是口语测试的目的必须明确；二是试题内容取样要有代表性；三是口语测试的内容对所有学生应该是公平的；四是测试的题目应该与测试的目的相适应；五是要使学生准确地理解测试题目的要求。

效度研究是一个关于有效性资料的积累过程，是通过积累证据对考试提供支持的过程 (Li & Li, 2014)。口语测试效度可以从不同的角度分为不同的种类，Zhang (2004) 把效度分为内容效度、效标关联效度和构想效度；Zou (2003) 把效度分为表面效度、内容效度、尺度关联效度和结构效度；Chai (2019) 认为效度发展主要经历过三个时期：分别为以相关分析为基础的单一效度观时期 (20 世纪 50 年代前)、分类效度观时期 (20 世纪 50 年代至 80 年代) 和效度一元化时期，效度种类也曾被分为构想效度、内容效度和效标关联效度，其中效标关联效度又分为共时效度和预测效度。

目前，关于口语测试的效度研究的文献大多是以构想效度为焦点。构想效度关心的是被试者在测试上的分数与我们根据能力或构想理论所做的预测之间的一致性程度。但由于目前口语考试的随机性和单一性，学习者的口试表现无法通过这种测试完全的体现出来，即现有的口语测试无法全面、有效地测试出学习者的学习状态和成果，因此这种测试无法达到教师前期预测的结果，这样往往效度就会很低 (Xu & Liang, 2020)。

Li & Li (2014) 通过对汉语口语考试 (Spoken Chinese Test, SCT) 的信度、试卷内部结构和共时效度的分析和考察，对 SCT 的有效性进行了认证研究。整个研究历经一年时间，通过收集各方面的数据，结果表明 SCT 这一作为计算机化的客观性口语考试形式的结果是稳定的、可信的。作者还通过将跨机构语言圆桌口语能力考试和汉语水平考

试口试 (HSKK) 作为外在效标, 发现三者成绩明显相关, 最终验证了 SCT 测试的有效性和可靠性。

Zhang (2019) 认为如今的汉语口语测试缺乏真实性和灵活性, 而口语测试应注重真实性与互动性, 尽最大程度还原交际的真实情景, 测得测试者的真实汉语口语水平。针对口语水平良好的考生而言, 在口语测试环节, 考官和测试者均要确保完成信息的接收与传递, 双方不停地变换听和说的角色, 因此测试者需尽量抓住考官所给出关键信息点, 并对要点进行详细地阐述, 从而取得优良的口语成绩。Zhang (2019) 通过对汉语教学中口语成绩测试测量的设计, 以流利性为例, 对测试中的相关指标进行筛选与分析, 得出流利性并不是一种主观评分标准, 而是可以通过统计的方式来得出具体得分。

关于口语测试的效度问题, Zhao (2011) 认为需要有一个框架来描述所测构想、实施作业以及对潜在能力的表现进行评估这三者之间的关系。她利用实验设计, 利用现有的实考录音, 对当时的 HSKK 高级的三种题型进行不同形式的组合, 通过研究不同组合形式下的评分结果与效标的相关性来探讨多种试题组合形式的效标关联效度问题, 结果表明实际考试 (评朗读和两个问答题) 和只评朗读和问答 2 均具有较高的效标关联效度。由此, 从提高效率的角度考虑, 只评朗读和问题 2 同样能够有效地评估考生的汉语口语水平。

针对效度在汉语口语测试不同阶段的表现, Li (2016) 研究发现效度在汉语口语测试的初级阶段的主要任务是制定出具体的测试所要达到的目的; 在汉语口语测试的中期阶段的体现在与如何让所出的口语测试试题最大限度达到测试的目的, 从而保证测试的效度; 在汉语口语测试的后期阶段主要体现包含在评分员的操作上, 受过有效专业培训的评分员在评分前学习评分标准, 有利于对口语测试的目标能力达成共识, 避免评分出现较大差距。

关于口语测试的效度研究是在不断发展的, 在传统的汉语口语测试的基础上, 为了进一步提高汉语口语测试的效度, 交互式语音任务成为研究的热点话题。互动能力是指参与者设法克服潜在语言障碍、实现交流目标的能力, 在分析中分为话语补偿、商讨、澄清和改正误解四类。Bachman & Palmer (2010) 通过自行设计的口语任务, 让测试者在模拟环境中完成基于真实交际目标的任务, 并以标准化的方式从互动中考察测试者口语水平在测验分数中的方差贡献, 从而验证其口语测试的效度。

综上所述, 效度定义的发展经历了从单一概念到种类概念, 再到整体概念的历程, 最终形成了科学、融合的观念。关于口语测试的效度研究虽然已经取得了不错的研究成果, 但国际中文教学领域仍然在针对“效度整体观”观念下的口语测试效度的内涵、外延和验证方法等方面的内容进行理论探索和实证研究, 继而推动了汉语口语测试效度研究向着纵深方向发展。

3.2 存在问题与后续改进方向

本研究在梳理近二十年来汉语口语测试的效度与信度研究文献的基础上,发现尽管该领域取得了一定的进步,但仍存在一些问题与挑战,这些问题的存在影响了汉语口语测试的科学性和可靠性。同时,基于现有研究的局限性,本部分也将提出可能的后续改进方向。

目前,汉语口语测试的效度问题主要体现在内容效度和构念效度上。内容效度方面,现有测试往往难以全面覆盖口语交际能力的各个方面,特别是在社会语言能力和策略运用能力的考察上不足。例如,许多测试更侧重于语法和词汇知识,而忽视了交际能力或互动技巧等关键要素,而这些是口语交际中不可或缺的部分(McNamara, 2000)。构念效度方面,部分测试未能充分反映出口语交际中的真实场景和任务,导致测试内容与实际交际能力之间存在偏差。这些测试可能过于简单或不真实,无法有效评估考生在实际复杂社交互动中的表现(Hu, 2016)。针对这些问题,未来研究可通过深入分析汉语口语交际的特点,设计更具代表性和挑战性的测试任务,以提高测试的内容效度和构念效度。

与此同时,随着技术的发展,汉语口语测试的方式和手段也在不断革新。Li & Li (2014)认为关于测试,每一个考试项目都是在不断累积支持效度证据的过程中走向成熟的,运用自动信息处理技术测量汉语口语水平是一次技术革新,且通过研究证明该考试可以在任何时间、任何地点通过电话或计算机进行,而且不需要人工进行评分,自动评分可即时生成客观、可靠的结果。

进一步地,跨文化交际能力的考察是汉语口语测试的另一个需要关注的方向。随着汉语国际推广的不断深入,考生来自不同的文化背景,如何在测试中合理评价他们的跨文化交际能力,是提高测试效度的重要课题。未来的研究可以结合跨文化交际理论,设计能够考察考生在不同文化交际情境中的适应能力和策略运用的口语测试任务。

汉语口语测试研究的多元化和国际化也是未来发展的趋势。目前,该领域的研究主要集中在中国和部分汉语学习者人数较多的国家,对其他地区和不同类型学习者的汉语口语测试研究相对较少。因此,扩大研究的地域和对象范围,进行跨文化和跨地区的比较研究,将有助于提升汉语口语测试效度研究的普遍性和适用性。

汉语口语测试的效度与信度研究虽已取得一定成果,但仍需在内容与构念效度、信度提升、技术应用、跨文化能力考察以及研究范围的拓展等方面进行更为深入的探索和改进。通过这些努力,可以有效提升汉语口语测试的整体质量,更好地服务于汉语作为第二语言的教学与评估。

4. 信度考量与质量控制

4.1 二十年间信度研究的主要发现

在过去二十年的研究中，信度作为汉语口语测试质量的重要指标，其研究成果丰富而深入。信度主要反映了测试结果的稳定性和一致性，是评价测试可靠性的核心指标。在此期间，多项研究致力于探究汉语口语测试的信度问题，并提出了相应的改进措施。

信度，也被称为可靠性，是语言测试质量评估的重要指标，是效度的前提。Zhou(2009)提出测试的信度是指它的结果对受试者的表现而言的可靠程度。一个好的语言测试必须具有较高的信度，误差越小，信度越高。信度和效度是既统一又对立的两个要素。一项好的测试，它必须具有较高的信度和效度。

口语测试的信度，主要取决于测试考官对考卷评分误差的控制，以及对其标准的掌握是否一致(Zhao, 2012)。无论是主观型还是客观型的口语测试，都包括命题、施测和评分三个基本过程(Kane, 2013)。由于口语测试是对被试者口语语言表达能力的考核，属于主观型测试，除了命题维度外，制约口语测试信度的还有施测维度和评分维度(Chai, 2003)。Zhang(2007)也认为高质量的考试应该尽可能减少命题、施测、评分过程引起的误差，确保被试的考试结果是考试要测量的语言能力的反映，否则考试的信度就不高。而在汉语口语测试中，命题、施测和评分过程对考试信度都有影响。

Kane(2013)指出，命题是语言测试的核心环节，其目的是通过合理设计的试题评估考生的语言能力。Yang(2009)提到，汉语口语测试题型应具备多样性，包括朗读、问答、看图说话、角色扮演等，这些题型能够全面评估考生的语言知识和交际能力。他还强调，题型的选择应基于语言水平和实际交际需求。无论是主观型还是客观型口语测试，命题是整个测试过程的基础环节，直接影响施测和评分的有效性和信度。

施测即测试方法、测试过程和测试程序能否确保被试者将其口语水平尽可能多地表现出来，也就是能否保证被试者的口语测试表现或获得的言语测试样本是被试者真实口语水平的反映。Chai(2003)认为对口试信度的影响主要体现在施测方法的选择是否科学，测试程序、考官因素等能否最大限度地保证测试者在口语考试过程中将其真实的口语水平表现出来。

所谓口试评分就是考官采用某一特定评分方法、根据相关固定的一套评分标准和评分细则、参照口语水平量表对口语测试表现或言语样本等级所做的描述、对测试者的口语水平作出主观判断并评分的过程。在汉语口语测试中，评分问题是研究者最关心、学术界最有争议的环节，主要是由于无论什么类型的测试，采用什么样的评分方法，对汉语学习者口语测试的最终评价都无法摆脱考官的主观判断误差的影响。而汉语口语测试

评分的误差主要源于三个方面：口语水平的多维不确定性、口语测试的答案难以客观量化和评分员主观判断的差异性（Zhang, 2007）。

口试评分质量的研究包括评分一致性研究和评分准确性研究两个方面。Chai(2003)从这两个方面对 HSKK 高级评分的信度进行了研究，结果表明 HSKK 高级的评分员信度、组间信度、历时性信度和分类一致性信度均比较高；评分的准确性和区分度还有提升的空间。

Zhang (2007) 认为，口语测试的信度不高通常表现为不同评分员之间、同一评分员在不同的心理和情绪状态下，对同一考生口语水平判断的不一致性。她主要采取了两种方法，一是请两名不同的口语老师对考生进行评分；二是三个月之后，请其中一名老师再次对考生的语言样本进行评分，以此来推断口试的信度。

随着汉语口语测试考生数量和考试规模的不断扩大，针对汉语口语水平测试高级（HSKK 高级）的评分信度，Zhao (2012) 运用现代测量理论，对高等汉语口语测试的三类题目（朗读、以叙事性和描述性为主的问答和以说明性和议论性为主的问答）及其不同组合形式的评分信度进行了考察。她认为口试的信度主要取决于阅卷者对评分误差的控制，以及对其标准的掌握是否一致，而概化理论以其在信度估计上的优势，成为主观型考试信度研究的主要理论。概化理论的分析结果显示，综合被试在三类题目上的表现给出的口试分数最为可靠；只评被试在议论性题目上的口语表达，同样具有很高的测试信度。

总体来说，命题质量、施测程序和评分误差的控制是口语测试的三个基本维度，其中评分误差的控制对语测试的信度影响最大。为了提高评分的一致性和准确性，提高口语测试的信度，必须对评分员、评分方式和评分形式对口语测试的评分结果的影响进行动态性和追溯性评估。

4.2 信度提升的措施及其效果评估

在汉语口语测试中，信度是衡量测试结果可靠性的重要指标，它反映了测试工具的稳定性和一致性。为了提升汉语口语测试的信度，并确保测试结果的可靠性，采取了一系列措施，并对其效果进行了评估。

首先，对测试评分者进行了规范化培训，以减少评分的主观性。评分者培训包括对评分标准的深入解读和评分实践的模拟练习。研究表明，经过系统培训的评分者在进行口语测试评分时的一致性显著提高，从而增强了测试的内部信度。例如，Wang (2016) 的研究显示，经过培训的评分者之间的克伦巴赫 α 系数从 0.70 提升至 0.85，反映了评分的高度一致性。

其次，优化了测试题目的设计，使口语测试题目更具有代表性和针对性。通过对历年汉语口语测试内容的分析，研究者对题目类型进行了调整，以覆盖更广泛的语言技能

和交际场景。另外，增加了测试的样本量，确保测试结果能够更好地泛化到总体。Gao 等（2013）的研究发现，经过题目优化后的口语测试能够更准确地反映学习者的口语交际能力，测试-重测试信度从 0.65 提高到 0.82。

再次，学者建议引入了多样化的评分方式提升口语测试的信度，如自动化评分系统和多评分者评分。自动化评分系统通过语音识别和自然语言处理技术为测试者的口语表现打分，减少了人为评分的偏差。引入多评分者共同评分的机制，通过统计学方法（如双评分者相关性分析）来监控和校正评分偏差。研究发现，这些措施显著提升了测试信度。例如，Zhou（2009）的研究指出，采用自动化评分系统后，汉语口语测试的信度从 0.75 提高至 0.90。

定期对测试进行信度分析和反馈机制的建立也是提升信度的重要措施。通过对测试结果的统计分析，识别信度低的题目，并进行修改或替换。建立了测试后反馈机制，收集测试者和评分者的意见，及时调整测试流程和评分标准。这些措施有助于不断优化测试工具，保持其信度在较高水平。Wang（2016）的研究证实了这一点，研究中的信度分析帮助研究者发现了测试中的若干问题，并据此进行了改进，信度提升了约 10%。

此外，信度问题也是汉语口语测试需要面临的挑战。当前，口语测试的评分往往受到主观因素的影响，如评分者的个人偏好和经验水平等，这在一定程度上削弱了测试结果的一致性和可重复性。为了提高信度，未来研究应加强评分者培训，建立更为严格和统一的评分准则，同时也可以采用多评分者同时评分的方式，以减少主观偏差的影响。

通过对评分者进行规范化培训、优化题目设计、引入自动化评分系统和多评分者评分机制以及建立信度分析和反馈机制等措施，汉语口语测试的信度得到了显著提升。这些措施的实施不仅提高了测试结果的可靠性，而且为后续的汉语口语测试研究和实践提供了重要的参考。未来研究应继续关注汉语口语测试信度的提升，特别是在不断变化的教育评估背景下，如何利用新技术和方法进一步优化评分过程，增强测试的可靠性和实用性。

5. 结语

本文通过系统性的文献综述，详细梳理了近二十年来汉语口语测试在效度与信度方面的研究动态与进展。研究表明，汉语口语测试在逐步建立综合性与交际性评价指标的同时，评分体系和标准也得到了相应的完善。但是，在评价过程的具体实施中，依然存在着评估准确性不足和一致性较低的问题，这些问题限制了汉语口语评测的发展。因此，汉语口语测试的效度与信度研究尚需深化，以促进其科学性和公正性的提高。

针对现有研究的不足，未来研究可以从以下几个方面深入展开：首先，加强对不同汉语学习者背景和能力水平的口语测试适用性研究，以保证测试的普适性和精准性。其次，创新技术手段的应用，如计算机辅助评分、人工智能等，可以用于辅助汉语口语测试，提高其效率和客观性。此外，应当增强跨文化交际能力的测试评价，在保持汉语本土文化特色的同时，更好地适应国际交流的需要。最后，为确保研究的深入和实用性，建议构建跨学科的研究团队，集合语言学、教育心理学、测试学等领域的专家共同推进汉语口语测试的效度与信度研究，以期更全面地服务于汉语作为第二语言教学的实践需求。通过以上建议的实施，期望能使汉语口语测试更加科学化、规范化，从而为汉语国际教育质量的提升做出积极贡献。

参考文献

- Bachman, L. F., & Palmer A.S. (2010). *Language Assessment in Practice* (2nd ed.). Oxford University Press.
- Chai, S. S. (2003). The theoretical analysis and experimental research of rater reliability of Chinese oral proficiency test. *Language Teaching and Linguistic Studies*, (04), 69-77. (In Chinese)
- Chai, S. S., & Wang, Y. Q. (2019). Review of validity research based on Journal of Language Testing during 2008 to 2018. *Examinations Research*, (05), 60-67. (In Chinese)
- Gao, H. Y., & Liu, F., & Ji, H. Q. (2013). Construct validity in language testing: Development contribution and challenges. *Journal of Northeast Normal University (Philosophy and Social Sciences)*, (05), 135-138. (In Chinese)
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2004). Coh-Metrix: Analysis of text on cohesion and language use. *Discourse Processes*, 37(2), 91-112.
- Hu, H. (2016). Rethinking Chinese oral proficiency assessment: A review of test design and validity. *Journal of Language Testing and Assessment*, 22(3), 35-49.
- Huang, Q., & Wang, H. (2019). Research on the reliability and validity of automatic oral test by machine. *Modern Educational Technology*, 29(09), 59-65. (In Chinese)
- Kane, M. (2013). Validating the interpretations and users of test scores. *Journal of Educational Measurement*, (50), 1-73.
- Li, X. Q., & Li, J. H. (2014). Validity analysis of spoken Chinese test. *Chinese Teaching in the World*, (01), 103-112. (In Chinese)

- Li, Y. H. (2016). The relationship between reliability and validity of language test. *Journal of Jiangsu Normal University (Philosophy and Social Sciences Edition)*, (05), 88-92. (In Chinese)
- Lu, X. (2014). Speech recognition technologies and their implications for language testing. *Language Testing*, 31(3), 345-367.
- McNamara, T. (2000). *Language Testing*. Oxford University Press.
- Phakiti, A. (2008). Construct validation of Bachman and Palmer's strategic competence model over time in EFL reading tests. *Language*, (25), 237-272.
- Wang, J. M. (2002). A study of the scoring of three types of oral test items. *Chinese Teaching in the World*, (04), 63-77+3. (In Chinese)
- Wang, T. Y. (2016). A study on the influencing factors of construct validity in language testing. *Course Education Research*, (23), 25-26. (In Chinese)
- Weir, C. J. (2005). *Language Testing and Validation: An Evidence Based Approach*. Palgrave: Macmillan.
- Xu, X. Y. (2005). On Chinese oral test. *Journal of Yunnan Normal University (Teaching & Studying Chinese as a Foreign Language Edition)*, (05), 39-45. (In Chinese)
- Xu, Y. C., & Liang, X. J. (2020). A research on the validity and reliability of the oral Chinese test for international students. *Journal of Changchun Education Institute*, (05), 29-35. (In Chinese)
- Yang, H. Y. (2009). *Fourteen Lectures on Teaching Listening and Speaking in Teaching Chinese as a Foreign Language*. Peking University Press.
- Yang, Q. (2020). *A Comparative Study of Chinese Oral Communication Strategies Between African Students and Japanese and Korean Students*. Shanghai Normal University. (In Chinese)
- Zhai, Y. (2012). On the evaluation criteria for Chinese oral achievement test. *Teaching Chinese to Speakers of Other Languages Studies*, (01), 42-50. (In Chinese)
- Zhang, F. F. (2007). Research on score reliability of spoken Chinese test. *Sun Yatsen University Forum*, (11), 175-176. (In Chinese)
- Zhang, K. (2004). Measurement is a component of a scientific theory: A restudy of construct validity. *Journal of Yunnan Normal University (Teaching & Studying Chinese as a Foreign Language Edition)*, 02(05), 39-45. (In Chinese)
- Zhang, Q. (2019). Research and reflection on the validity of oral tests in Chinese as a foreign language class. *Occupation*, (06), 86-87. (In Chinese)

- Zhao, Q. F. (2011). A study of the criterion-related validity of the item different combinations of spoken test of HSK (advanced). *Journal of China Examinations*, (09), 16-21. (In Chinese)
- Zhao, Q. F. (2012). On the reliability in spoken test of advanced Chinese proficiency test and implications for teaching. *Chinese Language Learning*, (05), 89-94. (In Chinese)
- Zhao, S. M. (1999). The discussion on the validity of oral proficiency test scores. *Journal of Mongolia Normal University (Philosophy and Social Sciences)*, 28(06), 125-129. (In Chinese)
- Zhou, Z. H. (2009). Reliability and validity of second language testing. *Journal of Language and Literature Studies*, (19), 154-155. (In Chinese)
- Zou, S. (2003). The reciprocal relationship between language curriculum and language test. *Foreign Language World*, 98(06), 71-78, (In Chinese)