国际汉语教学学报（2023）
第 04 卷 · 第 01 期 68-80 https://doi.org/10.46451/ijclt.20230106

基于效度论证模式的加拿大中文项目分班测试评估

王姝娇 *
麦吉尔大学，加拿大

李智
萨斯喀彻温大学，加拿大

孙敏
魁北克大学，加拿大

摘要
为外语学习者进行有效的分班是外语学习、教学和语言项目管理成功的重要环节。加拿大大学中文课程尽管在多元化学生群体环境下蓬勃发展，但针对这种特殊背景下的分班问题的研究却十分局限。作为一次实证性的探索，该研究以凯恩 (Kane) 的基于论证的效度验证理论为框架，采用顺序解释性的混合研究方法，通过网站信息检索、社交媒体、个人访谈等形式收集数据，依据效度论证模式为理论框架，评估当前加拿大多所大学中文课程的分班测试。分析结果表明，大多数中文项目（学分课程）分班程序所使用的多元评估方式（如笔试、面试和背景调查相结合等方式）有助于考试效度的提高。该研究还就一些具有普遍性意义的问题进行探讨，如是否应该针对不同水平层次学习者研发不同测试版本、如何更好地实现评分质量控制与大纲或参考标准保持一致，以及如何确保在线测试效度等问题。

关键词
分班测试，效度，论证模式，加拿大大学汉语教学

1 引言

分班测试，又称为安置性测试，旨在评估学习者现有的语言水平以确定其适合的学习课程。这是语言教学中非常重要的环节，同时也是外语学习、教学和项目管理成功的关键因素。因其重要性，分班测试逐渐成为研究热点，但针对海外中文项目的相关研究尚处起步阶段。加拿大汉语教学可追溯到二十世纪初一些神职人员与中国的互动。高等院校是海外汉语教学的重要阵地，如麦吉尔大学（McGill University）早在 1930 年代就已设立汉语语言项目，英属哥伦比亚大学（University of British Columbia）1957-1958 年首次开设汉语课程。时至今日，随着中国移民及汉语学习者的增加，汉语已成为加拿大继官方语言英语和法语之后使用最为广泛的语言（Statistics Canada，2017）。加拿大 96 所正规高校中至少有 40 余所大学开设了各类汉语课程，注册人数稳中有升，课程设置日趋多样，汉语项目蒸蒸日上。汉语教学的对象在语言水平、文

* 通讯作者。联系电邮：shujiao.wang@mcgill.ca

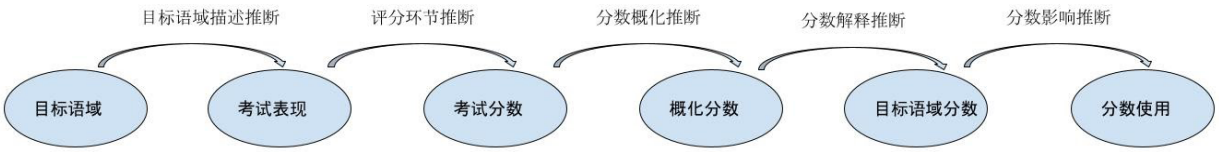
化背景、学习动机和专业需求等方面都表现出了极强的多元化与多样性特征（宋，2013）。有些大学针对不同的学生群体，分设华裔与非华裔、普通话与方言（如广东话）课程；针对学生的不同语言水平开设初、中、高等不同级别的课程；为满足学习者需求开设商务汉语、考试培训类（如 HSK）和汉语教学法等课程。因此，“如何分班”一直是加拿大开设中文课程的各所大学共同关注的热点和难点问题。复杂的学生背景对分班测试来说是一个巨大的挑战，而以往以笔试和面谈为主的测试方法所带来的额外工作量也对科学有效的分班测试提出了现实需求。基于上述原因，探索和研究准确的、有效的分班测试方式显得尤为重要。

2 文献回顾

2.1 效度验证

效度研究作为语言测试领域的核心问题，其指导框架经历了三个发展阶段，即 (1) 基于标准的效度模式; (2) 构念模式; (3) 整体效度模式。其发展过程体现了由传统“分类效度观”到“整体效度观“的整合和过渡。基于论证的效度理论框架（Kane, 1992, 2013）就是在这一背景下产生的，与“整体效度观”一脉相承。如图一所示，该框架认为研究者首先要考虑考试分数所代表的意义和使用分数产生的影响；然后对相应的分数以及分数使用进行论述和推断（inference），包括目标语域描述、评分环节、分数概化、分数解释（构念）、分数类推（extrapolation），分数使用及影响等；最后，根据收集到的相应的定性或定量数据，对考试分数相关的效度进行论证。此框架的理论与实践被广泛地运用到各类教育测试项目中，效度研究的中心也转移到解读分数意义及分数使用所产生的影响上。

图 1. 基于论证的效度研究框架



基于论证的效度研究理论框架使用的论证推理环环相扣并兼具方法论意义，可以指导研究者就目标语域、评分、分数解释和测试构念以及分数使用和相关决策造成的影响等方面开展研究及实施论证。知名的实证研究包括 Chapelle 等人 (2010，2012，2020)，他们应用该理论框架对托福考试开展了一系列细致的效度论证；Li (2015) 就美国高校英语作为第二语言的分班测试进行了效度验证及推理，为考试质量评估和改革奠定基础。早期关于效度理论及验证模式的研究主要集中于英语测试领域，国内文献主要以相关领域的发展的引介或述评为主（韩宝成、罗凯洲，2013）。近年来凯恩（Kane）的效验框架在语言测试领域不断传承和发展，针对不同语种考试的研究相继展开，也出现了较为著名的有“测试使用框架”（Bachman 和 Palmer, 2010）和“基于证据的效验框架”（Weir, 2005）。汉语作为第二语言的测试研究领域也参考其成果，将其应用于大型高风险考试的研究，如 Wang (2021) 基于论证模式针对 HSK 的结果效度，即 HSK 考试在课堂（微观）到社会领域（宏观）的影响方面进行验证。Bachman & Palmer 和 Weir 的框架从社会认知视角出发，更注重后果、决策和解释，但是其架构较为抽象，对于非高风险大型测试（如分班测试）的可操作性不易把握，况且测试开发的效果也有待进一步验证。

2.2 分班测试

分班测试因其独特的属性及对教学的指导意义一直以来备受研究者关注，其发展历史与语言测试理论的发展一脉相承。从二十世纪后期开始，为了满足各院校较大规模招生、快速分班的需求，外语领域（主要以英语为主）通常采用标准化考试作为分班的主要工具，但因其去语境化、以语法为中心、客观题型等特征与分班测试的目的不相匹配而被诟病（Kokhan, 2013）。此后经调整，以分班为目的的语言测试通常包含对以下内容或技能的考查: (a) 词汇，(b) 翻译 (母语翻译成其它目标语言)，(c) 阅读理解，(d) 自由写作，(e) 语法，(f) 听力理解，(g) 发音，和 (h) 口语表达等。

经梳理中外文献（例如 Bernhardt, Rivera & Kamil, 2004; Heilenman, 1983; O' Sullivan, 2011; Shohamy, 1998；魏妙纯，2020 等) 总结发现，合理有效的分班测试通常应具备以下特点：(1) 应对学生群体的异质性（heterogeneity）; (2) 强调语言能力表现；(3) 具有明确界定的成绩标准；(4) 具有针对最低合格程度的成就描述; (5) 配合项目课程的设置与安排;（6) 明确的教学目标; (7) 提供特殊的学习机会；(8) 实现本地化等。

过去十年，汉语分班测试的研究取得了长足的进展，但研究主要集中于国内高校针对留学生的分班策略研究和分班测试应用研究。柴省三（2011）总结了可用于分班测试目的测试，其中有标准化语言水平测试（HSK），教学机构自编书面测试，教学机构教师面试测验，学生语言水平自我评价等。在分班测试的简化测试形式方面，任春艳（2007）就基于汉语基本句式的习得顺序研究及语言能力测试理论，进行以汉语句式作为考查内容和分班指标的简化测试形式的实证研究。伍秋萍，洪炜，邓淑兰（2017）基于阅读理论模型，进行了将汉字认读作为建构简易汉语能力鉴别指标的实证研究。马存燕（2013）提出可以通过研发多套试题，建立测试题库，使学生在入学报名前通过网络测试平台自测了解自己的基本汉语水平。

针对分班测试结果的信度和效度，各位学者也进行了初步的实证探索。例如，辛平（2007）对笔试难度系数进行了探讨，提出分班笔试的难度系数应在 0.5 左右为宜，写作主观题在区分中级和高级水平中效果显著。李小萌（2008）运用 Bachman 的交际语言测试模式，尝试在题型的设计和开发上更为全面地考虑受试者的多个方面的语言交际能力，从而提高入学分班测试的效度。 柴省三（2011）提出了在汉语入学的分班测试的决策实践中运用多元聚类分析法提高分班测试分数的决策效度，确保课堂教学的针对性。罗莲（2012）在实证研究中基于某次分级测试的成绩，建立了判别分析模型，通过判别分析与教师评价的效度对比，证明了使用专家判定与判别分析相结合的方法有利于分班的决策效度。郭修敏（2017）最新一项实证研究中编制了分级测试客观试卷，结合听力和阅读两项技能对留学生进行分班，获得了较为理想的测试信效度。

海外国别化的分班测试研究为数不多且主要集中于个案研究，如宋刚（2013）对某孔子学院中文项目分班测试的讨论，Song (2008) 对美国某大学中文课程分班测试的研究。然而，这些研究强调分班测试个性化和灵活性，主要注重发现问题，但对目前分班测试难点（如准确性、可操作性、标准化等）存在分歧，解决方案的有效性有待进一步论证。

3. 研究方法

3.1 研究框架

本文以凯恩（Kane）基于论证的效度理论为研究框架，从分班测试结果的使用出发，对公开的分班测试工具分别从目标领域、评分环节、分数概化、分数构念、分数使用决策及影响等

方面收集证据或者提出效度研究建议。将 Kane 的效度验证框架用于中文分班测试研究中，其主要推断和假设如下表所示：

表 1. 应用于中文分班测试研究的效度研究框架

目标领域描述（针对考试目标、考试说明及考试题型设计）
假设 1：考试内容规范规定与考试测量目标一致
假设 2：考试内容领域与中文项目使用的课程标准或者相关语言标准一致
评分环节（针对考生答题情况及考试分数）
假设 3：客观题与主观题的评分规则制定得当
假设 4：原始分数转化为量表分数的模型与观察数据拟合
分数概化（针对考试分数及考试试题）
假设 5：客观题与主观题的评分具有较高的分数信度。试题是样本的有效单元
假设 6：概化内容领域与目标领域一致
分数解释 / 测试构念（针对考试试题及考试结果）
假设 7：考生分数能够反映考试要求考察的构念或结构内涵
分数使用与决策的影响（针对考试结果和影响）
假设 8：分班决策的制定客观公正
假设 9：分班决策有利于学生的汉语学习

3.2 研究问题

本研究的主要研究问题是：

1. 加拿大大学的中文项目如何对学生进行分班？有何特点及难点？
2. 现行的主要分班测试形式效度如何？即是否能够准确地根据学生语言技能进行分班，语言技能是否被准确测量，考试内容是否与考试目的联系紧密，及考试是否可信？

3.3 数据收集与分析

本研究所使用的数据主要来源于以下三个阶段。第一阶段，收集整理加拿大境内 36 所英语大学和 4 所法语大学中文项目相关网站所列出的分班信息（如程序、考试形式、考题内容）；第二阶段，通过电子邮件或社交媒体分别与 10 名加拿大高校中文分班测试负责人进行有关分班测试的自由访谈；第三阶段，有针对性地与 2 名中文项目负责人及其学生的半结构化个人访谈。

收集的中文项目囊括了加拿大所有省份，具体分布情况如下：7 所大学来自加拿大西海岸地区（不列颠哥伦比亚省），7 所来自草原三省（阿伯塔省、萨斯喀彻温省、曼尼托巴省），22 所来自加拿大东部地区（安大略省、魁北克省），以及 4 所来自大西洋地区的省份（新不伦瑞克省、新斯科舍省、爱德华王子岛省、纽芬兰和拉布拉多省）。在数据分析方面，对收集到的加拿大 40 所大学中文项目的分班测试信息进行描述性统计分析、与汉语师生访谈的质性内容进行主题分析（thematic analysis）。两名研究人员分别就各项目分班相关事项（如程序、题目格式、能力构念）进行编码，并做出相应的比对分析。这些经编码所得的描述性统计结果主要用于院校间的比较。具体来说，对访谈进行归纳分析以确定相关主题，从而了解分班测试的特质和效度，进一步探索这些分班实践对中文项目、教师和学习者的影响。

考虑到加拿大各中文项目的多样性，在选择自由访谈对象时我们对中文项目规模（大型、中型、小型）、学校类别（研究型大学、综合型大学）和学生构成（华裔学生较多、以本地学生为主）等方面进行综合考量，尽可能选取有代表性的中文项目负责人（或教师）。其中 80% 的访谈对象同时担任分班负责人独立管理分班工作，部分高校的任课教师也参与分班测试或协调工作。在第三阶段更为深入的访谈时我们则选取了两位中文项目负责人（为保护隐私信息以下简称为甲和乙），他们从事汉语教学工作多年，曾建设或任教于海内外多个汉语项目，他们不仅教学经验十分丰富，而且对加拿大高校中文项目也具有深刻的理解和体会。经他们介绍，他们的两位学生随后参与了访谈，学生丙是一名本地学生，目前选修其所在学校高级汉语课程，另一名学生丁是一名华裔，听说能力较强，曾选修中级汉语课程。

4 研究结果

4.1 分班测试特点及难点

在收集 40 所大学中文教育项目网页信息时，我们发现 17 所大学的中文项目没有提及中文分班测试信息，仅 23 所大学的中文项目提供了分班测试的基本信息。我们还发现各中文项目采用的分班方式各有千秋，但考试内容、结构、决策等方面具有相似性和可比性。一般而言，零起点的学生直接进入初级班，完成项目内课程的学生可直接注册下一等级课程，而有汉语背景或在其它院校修过中文课程的学生都需经过分班测试进入与其水平相当的中文语言班。由于学生背景的复杂性，单一的分班形式往往具有一定局限性，不能达到理想的分班目的。因此，综合使用多种测试形式确定分班结果成为加拿大高校中文项目最为普遍的分班方式，包括 1）背景信息或者语言使用情况调查（11 所高校）（如学生家庭使用语言情况、在汉语母语区生活年限、参与国内九年制义务教育情况、到达加拿大的年龄、选修过的汉语课程和使用过的教材、通过的 HSK 等级）；2）笔试（7 所高校）（如由各语言项目统一要求开发的综合语言水平测试，包括听、说、读、写等技能，考察汉字、词汇、语法、翻译、作文等）；3）面谈（12 所高校）（如通过在线会议、电话或面对面等形式问询学生学习背景，根据教材涉及的主题或特定情景对话，阅读短文等，侧重考察学生的发音、流利程度、词汇或句式表达难易等语言表达能力）。

在我们调查的中文项目网页中我们注意到仅有 3 所高校明确提到使用笔试作为分班测试，5 所高校把笔试与口试列在一起，作为一个选项。此外，这些中文项目的网站上都没有提供笔试样题以及分数解释等重要信息。在我们考察的 40 个中文项目中，背景信息加面试或者教师咨询的分班模式是使用最多的一种。这些特征与加拿大中文教学与研究学会的加拿大中文课分班参考的调查报告结果一致（加拿大中文教学与研究学会，2022）。

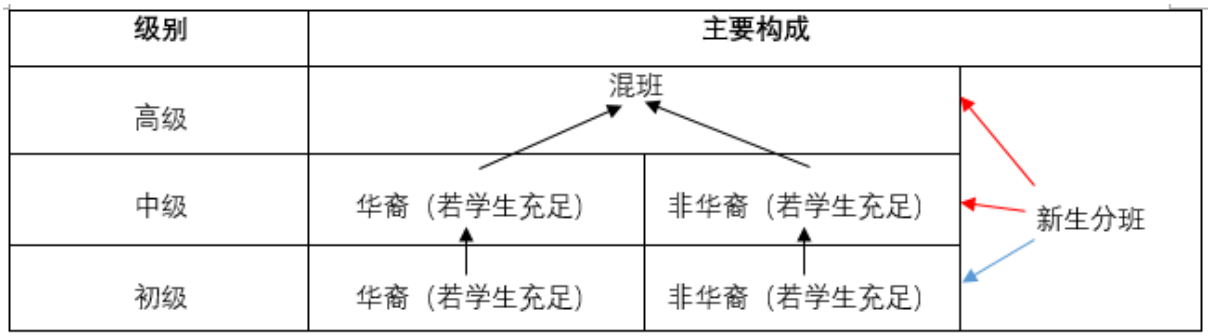
与加拿大各高校中文分班测试负责人的非正式讨论也验证了当前主要使用的分班方式以及在分班过程中面临的一些挑战。无论采用以上哪一种或多种考试形式，教师们比较关注学生的汉字掌握情况，除认读之外，汉字书写也是多个汉语项目分班的重要参考依据之一。在采访过程中，学生丙对此持有不同看法。他分享了一段过去参加分班测试的经历，他自我评估“在听、说方面水平不错，汉字认读和打字也没问题”，[1] 而在分班测试中因为汉字书写水平较低 [2] 而被分入低级班，随后任课教师发现其水平较高又重新调整了班级。他认为打字是未来发展趋势，即便在汉语水平考试 HSK 中也可以机考打字，汉字不应该成为外国学生学习中文的拦路虎。

此外，分班测试与项目本身课程设置的情况紧密相关。尽管加拿大各高校中文项目的规模不尽相同，仅就语言类课程方面基本设置为初、中、高等不同级别，但各校之间差别显著，同级别之间不能划等号。如位于加拿大东部城市的两所大学，以汉语三级为例，其中一所大学三

级[3]汉语课横跨整个学年，采用某高级汉语教材；而另一所大学的三级仅占一学期，学习内容也较为有限，仅为某中级课本的前半部分，因此对互转学分也带来挑战。总体而言，华裔学生较多的中文语言项目分班程序相对复杂，需根据学生的注册情况，设华裔和非华裔、普通话或方言班；根据学生的语言水平，开设不同级别的课程。相对而言，较小的中文项目或华裔学生较少的中文项目分班测试则较为简单，分班结果也较为理想。

经调查一些院校的分班流程和路径（如表 2 所示）基本为：零基础班通常不需要经过分班测试，直接进入；初级班（或低级班）会结合问卷情况（尤其是华裔学生）来判断；中、高年级由于人数相对少于初级班，所以通常采用混班的形式，如果是已选修过并通过考试的学生，则自动进入下一级，但针对新生（插班生）则需通过分班测试（或问卷、面试等）进入与之匹配的班级。

表 2. 部分高校总体分班流程（或路径）



从考试开发、实施、决策的整个过程来看，分班测试不仅对决策人专业知识（包括分班流程、考试内容、课程设置等）要求高，而且会增加他们的工作量，既耗时又耗力。这首先归因于多种形式的分班测试，但更多的是由于与学生通过电子邮件或一对一交流、协调（如商议考试时间、讨论考试结果等）带来的。尤其是有些中文项目规定全年都可以预约参加分班测试。此外，受人力和财力的限制分班负责人往往由中文项目负责人或任课教师担任，教师的工作节奏经常会被打乱。他们不仅要完成教学任务，而且要负责繁杂的管理工作。某研究型大学负责人表示，经常会遇到特殊的学生或学者，交换学生，学分转换等情况。各种情况因人而异，没有先例可供参考，都需要“特事特办”，操作起来非常麻烦。又如采访中某教师所言：“一方面如果情况允许，可以给负责分班工作的老师减少一部分授课任务，或为分班工作提供相应的资金；另一方面，采用与其他语言类课程相似的计算机分班测试（包含阅卷）都可能会对此情况有所改善”。

在谈到考试结果的准确性和公平性时，几位负责人一致认为大多数情况下分班测试的结果较为理想，但是总会出现一些“个例”。由于学生的程度个体差异较大，开设班级数量有限，他们发现把学生分到一个完全符合他们语言程度的班级并非易事，特别是中高年级或华裔背景的学生。教师会根据学生的具体情况推荐他们“就高”、或让他们等待适合的级别开班，或推荐他们注册“成人培训课程”。有些学生更倾向于选择低水平的班级以方便取得较好的成绩，因此他们有时会在分班测试过程中故意压低自己的汉语水平。而这类测试策略的使用都会增大分班决策的偏差。因此，在学生进入班级之后，都会有一个“缓冲阶段”用来调整和处理学生的语言程度和所分班级不符的问题。这需要由任课教师继续观察和评估学生的语言程度，及时发现语言程度和所分班级差异较大的学生，帮助他们协调换班。但由于时间紧迫，需要在开课的两周内做出决定，因此增加了难度。如果选课调整期限已过，则需学生、教师、项目负责人及行政人员等多方沟通协调才能进行换班。

在与各位项目负责人的交谈中，几位老师都提到同置于东亚系下的日语和韩语课程或同置于语言学系下的阿拉伯语和意大利语课程。与这些语言课程相比，中文项目不仅存在语言分班测试的不同，而且存在招生和课程设置等“竞争”，为了吸引更多的学生注册汉语课程，分班测试也会在某种程度上做出“妥协”。

4.2 分班测试效度

鉴于中文项目网页提供的分班测试信息以及分班测试的研究报告的局限性，我们将对比较普遍使用的背景问卷加访谈的分班测试模式进行分析，根据 Kane 基于论证的效度理论的研究框架提出相应的研究建议。

4.2.1 考试内容与目标

分班测试的内容通常与考试目标及课程标准一致。例如，在东部某大学中文项目的分班测试介绍中写道

> The sole purpose of this placement test is to provide an accurate picture of your knowledge of Mandarin Chinese as well as simplified Chinese characters. It is entirely about your actual knowledge of the language, not about the total score you can achieve on the test. [翻译：分班测试的主要目的是准确地反映您对汉语和简化汉字知识掌握情况。测试仅与您实际的语言知识有关，与您在测试中取得的成绩无关。]

测试包括自我评估，涉及听、说、读、写各方面，内容与形式可分为：1）与其课程与教材的话题联系紧密（如采用教材中的话题或生词）的考题类型；2）与课程无关（如具有普遍意义的话题）考题中主要以选择与判断等客观题型为主，但答案均增设“I don’t know the answer”[翻译：我不知道答案] 以避免” academic misconduct and misrepresentation with a negative impact on the validity and accuracy of your test score” [翻译：对你考试成绩的有效性和准确性有负面影响的学术不端行为和虚假陈述]。由此可见，考察内容集中于语言水平本身，内容规范与考试测量目标和课程标准都较为一致。但在访谈过程中，教师甲对这种较为常见的“万能测试“即根据同一试题、问卷或问题的结果来进行所有等级的评估表示担忧，“尽管这是比较方便和高效的测试方法，对老师评判也比较简单，但是在细分时（如区别高级和中高级）准确性不够高”。

4.2.2 测试评分与反馈

就分班测试的评分方面而言，网络公开的信息较为有限。除网络考试（客观题为主）自动评分之外，需要分班负责人或任课教师进行评分，体现出“灵活性”和“主观性”的特点。某负责人表示，不会根据每个题目的回答进行一一评分，而是综合各种测评方式的答题情况进行整体的评判。所以也没有什么统一的“评分标准”，主要看作答的情况也就是主要的印象，比如对汉字书写、发音、语法掌握和词汇运用各个方面进行一个综合评价。由此可见，此种评分方式是基于经验式的，尽管大多数学校参加分班测试的学生的作答情况或考试结果会记录备案，但主观类型的测试评分较为不透明。考试结果如果是客观题型的网络考试，通常在考试结束后立即显示考试成绩，但这一成绩并不代表分班结果。主观考试结果则不同，有时分班测试负责教师通过纠正表达的错误等和解释决策时对答题情况进行反馈，使学生了解大致答题情况。

4.2.3 考试结果与决策

与标准化语言考试不同，分班测试并不会根据听、说、读、写等能力和技能方面给出单项分数，教师在反馈时通常会提及语言知识（如生字、词汇、语法点）或某些技能（如语音语调、理解能力）等方面的回答情况，考试结果能够反映考试要求考察的结构内涵。在问及在决策时是否所有学生会一视同仁时，某负责人顿了一下，“也不能说完全一样，比如说有的学生一看就是装的，假装自己不会写汉字，要求去初级班，这种情况经常遇到……”。针对这种情况，许多高校的中文分班问卷或试题都会特别提及 Honour Code，例如：

> I certify that the information provided here is accurate and that I will fill out this placement test to the best of my ability. I also understand that, under XX University’ s Policy on Academic Integrity, providing false, incomplete or misleading information in this placement test is grounds for disciplinary measures ranging from a Disciplinary Notice to being required to withdraw from my program, or having my admission offer withdrawn by XX University. [翻译：我保证所提供的信息是准确的，保证我将尽我所能完成分班测试。我明白根据XX大学学术诚信政策，在本次分班测试中若提供虚假、不完整或误导性信息，会受到大学的警告、劝退或取消学籍等纪律处分。]

如果出现问题，学生本人需要承担相应后果。决策方面，也并没有统一的参考标准，但通常是根据分班测试的结果进行预分班。各校调研结果表明，绝大多数学生服从分班决策结果，但也有个别学生在开课初期换班或采用其它方法（如某学生的水平高于班级水平，但由于未开设更高级课程，可由老师相应调整对该学生的要求）找到适合的级别。因此，总体而言，决策的制定是相对客观公正的。

5 讨论

针对加拿大中文项目分班测试发现的难点和效度有待改进的问题，此部分将以表 1 为基础，围绕如何研发有效、可靠的分班测试，并分别从基于论证的效度研究框架中的各个推断部分进行展开。

5.1 目标语域描述推断

因为中文分班测试主要服务于本地的中文教学，其目标语域描述推断验证应该重点参考中文课程使用的教材以确保考试内容领域与中文项目使用的课程标准一致。马存燕（2013）认为预分班测试和分班测试应有别于通过特定的语言项目的测验来推测学生的汉语语言能力的水平测试，因为它们属于教学前测试，应该以学生的学历和教科书为重要依据。系统地描述目标语域可以帮助更好地理解本地的中文使用语域并为建立分班测试规范提供重要指导信息。但是考虑到学生群体的多样性，譬如“转学分”或在其他学校学过，有汉语基础的学生，以及参加过 HSK 考试的学生，我们建议各校根据自己的教材和课程规划，将学习目标与外在的相对独立的语言标准联立起来。这种做法在陈作宏与邓秀均 (2005) 的研究中也有提及。使用标准化语言水平测试试卷存在着可以使用的试卷有限、考试时间过长以及分数体系的解释复杂等问题，使用自编的试卷又常缺乏科学有效的研究。因此，通过参考客观的、统一的标准以实现等级标准与课程设置、教材内容及目标设定等方面的对接。例如加拿大不列颠哥伦比亚省 BCCAT（2020）的名为《中文核心能力框架与分班测试题库》（core competency frameworks and placement test

banks in the Chinese language）的文件对校际间互转学分就具有重要参考意义。此外最新中国教育部和国家语言文字工作委员（2021）发布的《国际中文教育中文水平等级标准》（以下简称《标准》）为海内外汉语教学提供了具有纲领性指导意义的参考标准，以《标准》为基准分别从知识（字、词、语法点等指标）、能力/技能（听、说、读、写、译）、话题任务等层面进行对接课程和考题，开发或对已有的试题进行重新归类、分级，设计评分标准兼顾所达等级描述（can do list）则可有效实现这一目的。鉴于用于不同标准之间的对接以及汉语测试方面的实证研究仍处起步阶段，这一方面的实证研究也将进一步提升理论对实践的指导意义。

在分班测试出题方面，内容基本涵盖听、说、读、写各项能力，但为了便于评分，某些分班测试，题目均以选择题为主，并且与目标语域、本地教学目标以及教材内容方面的联系相对“模糊”。Ji(2021) 通过对各语种继承语分班测试的文献回顾，提出了针对中文项目分班考题设计应注重接受型（receptive）与产出型（productive）并重，真实性（authenticity）与互动性（interactiveness）相结合的方式。在设计考题时，可考虑应用英语、艺术类课程中较为常用的“反向设计”（backward design）方法，从中文项目的总体目标出发（包括听、说、读、写、汉字书写等整体要求），结合各个级别、班级的教学目标（包括华裔班和非华裔班），同时参考或融入各级别、班级的考试题目、作业、作文等进行设计。试题结合教学目标有益于难度的把握，难度大不仅会导致对低水平学生的区分缺少准确性，而且会导致猜测率的提高。此外，不同等级的考试在选择题型时也应结合中文作为第二语言学习的特点，例如辛平（2007）对国内高校 170 多名留学生汉语分班测试结果的追踪研究发现，作为产出型的作文考试在对高水平学生的区分上具有不可替代的作用，但对中低程度的学生鉴别力相对较弱。建议在面向多层次的考生群体时，可考虑将作文测试作为附加考试，用来区分高水平考生，同时控制作文评分过程，减少评分误差。

5.2 施测与评分环节推断

目标语域描述推断确立之后的成果应该包括分班测试命题规范以及与目标语域相关的具体考试形式。从我们的研究来看，虽然各高校已经有了自己的分班体系（包括测试内容、形式、流程等），但针对不同测试（如背景调查问卷，面试和笔试）进行评分的客观标准则相对缺乏，因此，这也是中文分班测试研发与使用中面临的难题之一。

以背景问卷调查为例，它在分班过程中，尤其是对已有一定汉语基础的学生（例如华裔学生），发挥了重要的作用。Ji(2021) 对继承语学生的背景调查中较为有效的题目[4]和低效题目（如父母受教育程度、美国继承语教育的数量）进行了总结。如何对问卷题目进行评分，则因“校“而异，甚至因”师“而异，缺乏客观的、具有通识意义的标准。在规模较大、等级较多、华裔学生比例高的学校，仅依靠背景问卷调查测试，其结果的有效性相对有限。Thompson（2015）发现仅通过背景问卷方式分班有三分之一的继承语学生被分配到不合适的班级；此外，Liu（2011）的文章提到部分继承语学生为了避免被分到继承语班在回答问卷过程中可能会提供虚假信息。如何对背景调查问卷的信息评分尚且缺乏相关的实证研究，因此对该方面进行研究也是十分必要的。

新冠大流行开始后，面试方式由传统的面对面转为线上，带来了便利的同时也给测试评分带来了挑战。值得注意的是，有些因素可能会影响教师客观的判断，比如语音、语调等，不应以学生的发音能力评价其语言程度；同样，也不应根据书写汉字和使用汉字能力来区分学生的程度。某些学生可能读写能力差，但听说强。在网络面试过程中要求学生输入一段汉字来了解汉字掌握及书写能力正成为新的趋势。

对基于笔试的分班模式而言，中文项目负责人则需要考虑是否使用每级一卷，即为每个等

级、年级（除零起点外）各出一份试卷，或者采用“一卷多级”，即通过一张试卷根据作答情况评判学生水平。有能力开发基于计算机考试（机考）的中文项目可以考虑扩大并校准题库以帮助实现计算机自适应测试。

值得关注的是，一些学校的分班测试基于自我测评，即学生对自己的中文水平在各技能领域进行评分，学生通过标准化测试（及级别）或某选修课程、根据各等级的课程描述自我做出判断。自我测评是教育领域里的重要教学和学习工具，因其对学习者学习意识和学习成果的积极影响而得到广泛认可（Luoma，2013）。在数据收集和可获取性（accessibility）方面，与教师评估和同行评估相比，学生自我评估更易于使用，既方便又省时（Coronado-Aliegro，2006）。如 Powers 等人所述（2009），在参与者没有动机、故意歪曲他们的报告的情况下，自我评估往往更有效。

5.3. 分数概化与解释推断

在分班测试工具及施测定型后，中文项目或分班负责人应定期检查测试的分数信度，例如在使用以选择题为主的纸笔测试时应该计算并报告克隆巴赫（信度）系数（Cronbach’s alpha）或者其它类似的测试分数信度指标。在产出性技能测试（productive skills）方面，如面试或者写作题目，应定期开展评分人培训和评分校准工作，并检查和报告评分人之间的评分信度或者评分人内信度。相关的信度指标可以是评分一致性百分率或者是基于概化理论的概化系数。

在将考试结果和决策反馈给学生时，应结合考试目标，给学生提供较为详细、公正的评价（书面或口头），并记录在案。合理解释分班决策将有益于学生了解自己的真实语言水平以及与之匹配的课程设置，为日后规划语言学习奠定坚实的基础（Kim, 2015）。

5.4 分数使用与决策的影响推断

我们的研究发现，与其它语言的分班测试不尽相同，中文项目的分班测试成绩（分数）为分班决策服务，但并不起决定性的作用，分班负责人通常需要结合学生其他方面如背景问卷调查、面试等情况进行综合考量。如储诚志（2021）提到的在理想的分班测试过程中，学生个体的学习需要得到充分关照，也包括身体条件、学习条件、文化背景、学习兴趣等各个方面；与此同时，这些需求也应与课程设置及教学安排相吻合（如学校的政策和资源限制，班级人数要求，设备、教材、教师资源等）。此外，分班测试也应考虑外部效标，如学生的学习潜力、动机、性格等，应具有成长性。李海燕等人（2003）认为分班测试的目的之一是了解受试者是否具备在某个班级中学习的能力倾向，因此在衡量学生现有语言水平的同时，也应对学生的潜在学习能力有所关注。针对不完全符合分班条件的学生，教师也可因材施教，通过不同的教学策略帮助学生学习和成长，例如对于高级班的学生使用档案袋方法（portfolio），根据每个学生在学习课程期间取得进步的程度给出分数。对学生而言，准确的分班测试结果以及公正的分班决策也将对他们的汉语学习产生积极的影响，从主观上帮助其摆脱“学术不端”行为的动机。如若形成一个健康的生态环境，也将对教学以及整个中文教学项目形成正面影响。同时，分班负责人或教研室建立有效的测试后调整机制，每隔一段时间整理和分析分班测试结果，及时调整和改进考试内部效标，增加或完善题库，这样既有助于提高试题的质量，又有助于提高学术对分班结果的认可度。这一系列质量改进行为包括但不限于分析各题的区分度，调整试题使之适合课程发展的需要，关注语言测试和教育测量技术发展，考虑将新技术和成果应用于分班测试，使之具有长久生命力。

6 结语

随着汉语国际地位的不断提高、学习人数的日益增加，加拿大高校汉语项目面临机遇与挑战，设计并实施合理有效的分班测试是众多汉语项目亟需解决的重点和难点问题。本文采用定性研究和定量研究相结合的混合研究方法，基于 Kane 的效度论证框架全面系统地解析和评估了现有分班测试，并针对目前存在的问题和难点展开讨论。本文旨在通过对现行分班测评的反思，进一步优化分班程序、完善分班形式、改进考题质量，实现分班测试与目标设定、课程设置、教材内容、目标设定等方面的对接。由于本文调查对象仅基于部分院校的中文项目（尤其第三阶段仅为两校个案研究），因此在结论或建议方面也具有一定的局限性。下一阶段，拟通过语料库分析来挖掘加拿大中文项目中主流教材与《标准》指标的对比，通过对学习者各种测评成绩的相关性分析（correlation analysis），实现历时性追踪研究，确保我们了解分班测试对学习者在不同的学习阶段的影响，从而进一步推动海外汉语教学与测评的规范化与标准化。

注释

1. 作者注：该学生已通过 HSK 中级水平考试。
2. 作者注：据学生丙所述，体现在作文部分。
3. 该大学汉语语言项目共分为四个等级，从低到高依次为一级、二级、三级与四级，每个级别课程横跨秋、冬两个学期。
4. 主要包括（1）学生与继承语的接触，（2）学生对继承语的利用，包括频率、领域和环境，（3）在讲继承语的国家受教育的程度，（4）继承语教学的时间安排；（5）父母对继承语和文化的态度，以及（6）其他干扰因素，例如学生与当地官方语言（如英语、法语）的接触。

参考文献

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice.* Oxford University Press.

British Columbia Council on Admissions & Transfer. (2020). *Core competency frameworks and placement test banks in the Chinese language*. from https://www.bccat.ca/pubs/Reports/CoreCompChin2020.pdf

Bernhardt, E. B., Rivera, R. J., & Kamil, M. L. (2004). The practicality and efficiency of web-based placement testing for college-level language programs. *Foreign Language Annals, 37*(3), 356-366.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Chapelle, C.A. (2012). Validity argument for language assessment: The framework is simple. *Language Testing, 29*(1), 19-27.

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Sage Publishing.

Coronado-Aliegro, J. (2006). The effect of self-assessment on the self-efficacy of students studying Spanish as a foreign language. Unpublished Doctoral dissertation, University of Pittsburgh, Pennsylvania, USA.

Heilenman, L. K. (1983). The use of a cloze procedure in foreign language placement. *The Modern Language Journal, 67*(2), 121-126.

Ji, Jingjing. (2021). College-level placement for heritage language learners. *Foreign Language Annals, 54*(3), 690-713.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1):1-73.

Kim, A.-Y. (Alicia). (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*(2), 227–258.

Kokhan, K. (2013). An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing, 30*(4), 467–489.

Li, Z. (2015). *An argument-based validation study of the English Placement Test (EPT): Focusing on the inferences of extrapolation and ramification.* Unpublished thesis. Iowa State University.

Liu, J. (2011). Placement Test Development for Chinese Heritage Language Learners. *Journal of the National Council of Less Commonly Taught Languages,10,* 169-192.

Luoma, S. (2013). Self-Assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-5). Blackwell. https://doi.org/10.1002/9781405198431.wbeal1060

O'Sullivan, B. (2011). *Language testing: Theories and practices.* Palgrave Macmillan.

Powers, D. E., Kim, H.-J., Yu, F., Weng, V. Z., & VanWinkle, W. (2009). *The TOEIC® speaking and writing tests: Relations to test-taker perceptions of proficiency in English* (No. ETS RR-09-18). Princeton, New Jersey.

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*(4), 331-45.

Song, J. (2008). Developing placement instrument to meet the need of the diverse student population at university level - A case study of Chinese placement test at University of Hawaii. In Hudson, T. & Clark, M.(Eds.). *Case studies in foreign language placement: Practices and possibilities* (pp.119-131). National Foreign Language Resource Center at University of Hawaii, USA.

Statistics Canada. (2017). 2016 Census Topic: Language. https://www12.statcan.gc.ca/census-recensement/2016/rt-td/lang-eng.cfm

Thompson, G. L. (2015). Understanding the heritage language student: Proficiency and placement. *Journal of Hispanic Higher Education, 14*(1), 82– 96.

Wang, S. (2021). *Consequential validity of the Chinese Proficiency Test (HSK) from macro and micro perspectives*. Cambridge Scholars Publishing.

Weir, C. J. (2005). *Language testing and validation*. Palgrave McMillan.

储诚志（2021）我们该到哪儿去，美加中文项目的分班系统、对策与挑战，CLTA-SIG 华裔中文教学小组线上讲座。

柴省三（2011）关于留学生汉语入学分班测试决策效度的思考，《中国考试》，10：31-37。

陈作宏、邓秀均（2005）外国留学生汉语进修班分班测试初探，《云南师范大学学报（对外汉语教学与研究版）》, 5：32–38。

郭修敏（2017）面向 TCSL 的分级测试客观卷开发实证研究，《世界汉语教学》，31：242-252。

韩宝成、罗凯洲（2013）语言测试效度及其验证模式的嬗变，《外语教学与研究》，3：411-425。

加拿大中文教学与研究学会（2022）加拿大大学中文课分班参考。Retrieved from https://sites.google.com/site/cltraassociation/uu/chinese-language-placement-workshop-in-canadian-universities/test

李海燕、蔡云凌、刘颂浩（2003）口语分班测试题型研究，《世界汉语教学》，4：79-89。

李小萌（2008）试析交际语言测试模式在入学分班测试中的应用，《科技资讯》，23：223-227。

罗莲（2012）汉语作为第二语言的分级测试题型研究，《语言教学与研究》，2：9-16。
马存燕（2013）短期速成汉语学习的预分班测试模式分析，《现代语文（语言研究版）》，4：159-160。
任春艳（2007）关于简化分班测试的实验研究，《语言教学与研究》，6：45-50。
宋刚（2013）海外华文教学中的分班测试，《海外华文教育》，4：420-427。
伍秋萍、洪炜、邓淑兰（2017）汉字认读在汉语二语者入学分班测试中的应用——建构简易汉语能力鉴别指标的实证研究，《世界汉语教学》，31：395-411。
魏妙纯（2020）美国大学中文分班方式，《华语文教学研究》，1：61-91。
辛平（2007）安置性测试的跟踪研究，《汉语学习》，6：76-81。
中华人民共和国教育部、国家语言文字工作委员 (2021) 国际中文教育中文水平等级标准 . http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202103/W020210329527301787356.pdf

投稿：2022 年 2 月 7 日；接受：2022 年 6 月 15 日；出版：2022 年 12 月 30 日

作者简介

王姝娇，博士，加拿大麦吉尔大学，测评专家，研究领域为教育测量与评估，语言测试。

李智，博士，加拿大萨斯喀彻温大学，助理教授，研究领域为语言测试、语料库语言学、计算机语言学。

孙敏，博士，加拿大魁北克大学，教授，研究领域为语言学、翻译学、汉语教学。

International Journal of Chinese Language Teaching (2023)
Vol. 4 (1) 68-80 https://doi.org/10.46451/ijclt.20230106

An Evaluation of the Placement Tests of the Chinese Programs in Canadian Universities Based on Kane's Validity Framework

Shujiao Wang
McGill University, Canada

Zhi Li
University of Saskatchewan, Canada

Min Sun
Université du Québec à Montréal, Canada

Abstract
Effective learner placement is critical for the success of foreign language learning, teaching, and administration in language programs. While Canada has seen a rapid growth of Chinese programs in Canadian universities with diverse student population, there are limited studies on the issue of placement tests in this specific context. As an empirical attempt, this sequential explanatory mixed methods study based on Kane's validity framework, aims to evaluate the current placement tests of Chinese programs in Canadian universities by employing information retrieval, social media, and semi-structured interviews. The findings revealed that most of the sampled Chinese programs (credit courses) claim to have placement procedures in use, which often involve multi-source evaluation (such as written exam, oral interview, and background questionnaire). This multi-source method is beneficial for improving the test validity. Common reported concerns in placement procedure were also discussed, including lack of test versions for learners of different proficiency levels or language status, the need for better rating quality control, alignment with agreed references or standards, validity of online test.

Keywords
Placement test, validity, argument-based, Chinese education in Canadian universities

Dr. Shujiao Wang, McGill University, Assessment Specialist, Research area: Educational Assessment and Measurement, Language Testing.

Dr. Zhi Li, University of Saskatchewan, Assistant Professor, Research area: Language Testing, corpus linguistics, computational linguistics.

Dr. Min Sun, Université du Québec à Montréal, Professor, Research area: Linguistics, Translation Studies, Teaching Chinese as a Second Language.