中文词汇多阶段自适应测试的构建及探讨

陈明君 * 杨宝玲 林同飞

摘要

个案机构在中文课程的设置上打破了一、二语相对独立的现状，以语言和语文连续统（Language and Literacy Continua）的概念重新整合中文能力，构建四个阶段、八个级别的能力框架，同时服务于中文一、二语的中小学生。为匹配该能力框架的落实与发展，个案机构以中文词汇作为标准考试的载体，采用专家判断和项目分析法，不仅建设了中文词汇多阶段测试题库，而且获取了每一道题目的难度、信度、效度及区分度参数。同时，为提高考试组织的效率和测试的信效度，为学生提供更为个性化的定制式考试，个案机构采用了计算机多阶段自适应模式（multistage testing，即 MST）施测，即不同中文能力水平的考生将因其作答情况获得不同难度值的题组。本文首先详述题库的建设，其次介绍如何借助题目参数开发多阶段自适应考试系统的过程，最后结合个案机构施测结果和教师访谈记录发现，中文词汇多阶段自适应测试在用时缩短、题目减少的情况下仍能较准确地估计学生的中文词汇等级。

*关键词：*词汇 多阶段 题库 自适应测试

* 陈明君，耀中耀华教育网络，联络电邮：249550734@qq.com (本文通訊作者)
杨宝玲，耀中耀华教育网络，联络电邮：yangbaoling@ycyw.cn
林同飞，耀中耀华教育网络，联络电邮：tungfei.lam@ycef.com

一、引言

国际学校的中小学汉语课程多分为中文作为一语（语文）、中文作为二语或 / 和外语（语言）几种。由于语言、语文课程相对独立，缺乏系统的整合，导致了课程目标和评估的范畴及顺序不衔接，经常出现课程和学习者错配或前者无法满足后者学习需要的现象。

语言和语文看似割裂，其实有着天然的联系。语言是语文的原型，而语文是一种建基于语言的沟通活动（a language-based semiosis）（Kress, 1997）。这种以发展的视角看待语言学习的观点在现代语言教育中是普遍存在的，不仅反映在各个国家和地区语言、语文学习标准的改革和优化中，也出现在国际文凭小学项目《语言范畴与序列》（Language Scope and Sequence）中。

本研究的个案教育机构（下文简称“个案机构”）在汉语课程设置中借鉴了语言和语文连续统（Language and Literacy Continua）的概念，于 2020 年 6 月至 2021 年 7 月重新设计汉语课程，梳理相关的语言能力范畴，最终形成适用于中文一、二语学习者及小学、初中学习阶段的四阶八级发展框架（林同飞等，2023）。该四阶八级发展顺序分别从字词、听说、阅读、写作等范畴，对不同学习阶段做了要求，教师可根据学生的具体情况，设置学习起点和终点，并根据学生的学习速度随时做出调整。

个案机构在汉语课程的目标和能力培养上做了有意的分级与衔接，与之相适配的学习内容和评估自然也要做出调整，尤其是每个学年初的分班测试，如何为学生提供多阶段、多级别的测评，辅助教师快速而有效地对学习者进行定级，为后续的差异化教学提供确凿的数据，值得探讨。

二、问题提出

学习者的中文能力涉及字、词、句、篇等各项语言要素，也涉及听、说、读、写各项语言技能。在语言要素中，词语是语言系统中最活跃、最有生命力的元素，因为语音依附于它，语法也需要在词语的

具体运用中体现出来，可以说在语言教学中词语有着无可取代的价值和意义。

词汇与语言能力之间呈现的密切相关性，已被众多学者所验证。词汇知识不仅和语言综合水平显著相关（Schmitt, 2010; Meara&Jones, 1988; Nation, 2001; Laufer, 1992），和听、说、读、写各项能力亦相辅相成（Schmit, 2010; Nation, 2001）。词汇知识的增长可以促进四种技能的提高，反过来，听说读写各能力的提高又有利于词汇知识的拓展（Nation, 2001）。Alderson（2005）对各种词汇测试（如词义、搭配等）分数和语言部分的测试分数进行比较，系统地解释了词汇知识和语言能力之关系。因此，以词汇作为中文测试的载体，以词汇测试的结果推测学生整体的中文水平应该是可行的。

如前所述，个案机构形成了四个阶段八个级别的中文能力框架，若以每一级别的词语组卷，将会形成八个级别的卷子。传统的多级别测试，正是一个级别对应一套或几套卷子。教师依据主观经验对学生的中文能力进行大致判断，而后提供学生相应级别的题目。无论学生能力如何，都需完成所有内容。如果学生中文水平和所选等级不符，则需要再次参加考试，才可能准确定级。这种考试模式虽然简单易行，但用时长，且准确性也有待商榷。

为了提高考试组织的效率和测试的信效度，个案机构采用计算机自适应测验（computer adaptive testing，简称 CAT），以期为学生提供更个性化的定制式考试。自适应测验是一种智能化测验形式，它能够根据学生前期回答问题的情况，选择与其能力最匹配的题目，然后根据学生当下给出的答案对其能力再次进行评估，之后再选出与他能力最匹配的题目供其作答，如此循环反复，直到满足测量准确或者长度的要求（郑蝉金、汪腾，2021）。

目前，计算机化自适应测试的模式主要有两种：基于题目层面的 CAT 模式（computerized adaptive testing）和基于题组层面的 MST 模式（multistage testing）（杨志明，夏胜俊，2021）。

CAT 模式以现代测量理论——项目反应理论（Item Response

Theory，即 IRT）为基础，通过建立被试者对测试项目的反应与其潜在能力特质之间的数学模型，实现对被试者能力的推测（刘洪峰等，2016）。CAT 模式结合项目反应理论和计算机技术，不仅能用更少的题目更快测量出考生的能力，而且能良好估计极端能力水平(Lord,2012; Wainer et al., 1992)，也就说纸笔测试中无法回避的“地板效应”和“天花板效应”在 CAT 里都可以得到相应的解决，因此得到广泛认可和实际应用。但 CAT 本身固有的特点导致其在考试实践中遇到了极大的困难（杨志明，2016；李贵玉等，2017）。如，大量区分度高的试题被反复选用，而区分度低的试题却很少被选用，导致前者过度曝光，而后者则浪费了题库资源（Chang,H.H. & Ying, Z., 1999）。又如，CAT 考试过程中不可修改答案，因为一旦修改了答案就会给能力估算带来问题。虽然一时的失误通过后续的做题也仍能回到适配的题目难度上，也仍能估算出学生的真实能力，但不可修改这一点与真实环境下的测试是相悖的，增加了考生的焦虑与压力。

为了弥补 CAT 模式的部分不足，计算机多阶段自适应测试（MST）应运而生。作为纸笔测试和 CAT 模式的“折中”（李贵玉等，2017），MST 是将一次测试拆分为两个或两个以上相互关联的阶段，考生在不同阶段被要求作答难度与其能力水平相当的微测验或题组（杨志明，2016）。相对于传统的固定顺序的纸笔测试，MST 可以根据被测试者的表现调整测试难度，提高测试的准确性和时效性。和 CAT 相比，它可以更好地管理测验、控制题目曝光度、做等化[2]处理和题目扩充等。MST 在国际主流的教育评估中已被广泛使用，如 PISA（Programme for International Student Assessment）和 PIAAC（Programme for the International Assessment of Adult Competencies），并获得了良好的效果。

本研究即以词汇作为考试载体，建设多阶段自适应测试系统，具体研究问题包括：

2 等化（equating）是指使用统计方法，将一测验的分数转换至另一测验分数量尺，以比较两个测验分数关系的过程，其目的是为了校准试题难度的差异。（蓝珮君，2008）

（1）如何构建以词汇为本的中文多阶段测试题库？

（2）如何构建多阶段自适应测试系统？

三、研究方法

本研究采用混合研究法，具体包括专家判断法、项目分析和访谈，研究问题与研究方法之对应关系可见表 1。其中专家判断法用于建设有质量的题库。项目分析法用以获取每一道试题的信效度、区分度和难度参数，验证题目质量的同时也为多阶段自适应测试系统的构建做好算法基础。访谈则辅助判定该测试系统的有效性以及作为后续系统改良升级的重要参考。

表 1：研究问题与研究方法对应关系

研究问题	研究方法
如何构建以词汇为本的中文多阶段测试题库？	专家判断法
如何构建多阶段自适应测试系统？	项目分析法，访谈

四、研制过程

（一）第一阶段（2021.03–2021.08），词汇多阶段测试题库的构建

为支持大规模测试的组织，也为了测评结果的统计与分析，本研究采用选择题做为题型。编制选择题需同时考虑题干和选项。由于文本信息易对考生造成干扰（Schmitt, 2010），亦考虑到心理词汇的影响（薛琳，2019），因此题干以单句、图片和声音的形式呈现。选项的择取同样影响着题目的质量和有效性。本研究在黄山、石井秀宗（2022）梳理的八种常见词汇错误类型基础上，参考 Bloom 的认知分类（Testa et al., 2018）和相关的对外汉语偏误研究（肖奚强等，2020；刘祥友等，2012），拓展至十类常见错误类型，包括：音近、形近、反义、近义、同类、多义、同语素、搭配、最优和学术（见附录一）。在设计干扰项时，参考以上错误类型，做好备注，以达到题目错误类型的

平衡，也便于日后对题目做调整和分析。

除了题干形式、选项，测量的内容、题干的表述、题目的编排等也都影响着题目的质量。因此，本研究参照 Haladyna 和 Rodriguez（2013）的题目编写原则以及个案机构自行研发的教学分级词表，由三名研究员分三阶段参与题目编制和审阅。三名研究员中，两位具夯实研究背景，并有相关领域的研究成果，一位于个案机构从事实践工作十年以上，经验十分丰富。第一阶段，两名拥有学术背景的研究员参照选择题编写原则，从教学分级词表中分别挑选各级词语，编写四套子测试集 A、B、C、D。其中 A 测试集包含一级和二级词汇题目，B 测试集包含三级和四级词汇题目，以此类推。此后，子测试集初稿由这两名研究员进行交叉验证，判断依据仍为选择题编写原则，逐一核对每个题目的题干、选项、表述等等。对有疑问的题目进行标记，共同讨论直至意见达成一致。修订后的子测试集进入第二阶段。在第二阶段，同为此两名研究员，需各自独立地标注题目属性，标注的内容包括正确选项的词性、干扰项的错误类型、用途等。经过第一轮的标注，两名研究员进行交叉验证，同样筛出有不同意见的内容，互相讨论，以达成一致。若仍有意见分歧，将纳入第三名具丰富实践经验的研究员。这三者协商后，对标注信息进行调整，确认。第三阶段，邀请题目编写组之外的资深中文教师对题目进行试测和反馈，反馈的内容涵盖题目的清晰度、可理解性、梯度等等。组员根据反馈意见做最后的修订。至此，中文词汇题库已形成，八个级别，四个子测试集，共计 204 道题目。

（二）第二阶段（2021.09–2022.08），多阶段自适应测试系统的构建

多阶段自适应测试系统的构建共经过三步，分别是“题目参数的获取”，“多阶段自适应系统框架的搭建”和“多阶段自适应系统的部署”。测试系统完成构建后在个案机构展开大规模施测，本研究对施测后的数据进行了分析和说明。

1. 题目参数的获取

第一阶段所编写的四个子测试集于2021年9月份进行施测，共收取作答数据4464份。作答数据使用Ishii和Huang（2021）所开发的系统进行分析，获取题目信效度数值。IRT计算和等值工作则使用“Irtoys系统”（Partchev等，2017）。分析结果显示：四个子测试的Cronbach系数是0.92，0.93，0.95和0.89；取其中某一校区的阅读测试[3]成绩作为参照，四个子测试集的效标关联效度结果为A 0.54（n=128），B 0.67（n=62），C 0.72（n=148），D 0.69（n=67）；中文词汇测试整体的区分度为1.75，不同子测试集的区分度没有显著差异；词汇测试的整体难度为1.29，每个子测试集的难度依次递增。这说明题目在信度、效度、区分度和难度上达到统计学标准（黄山、石井秀宗，2022）。

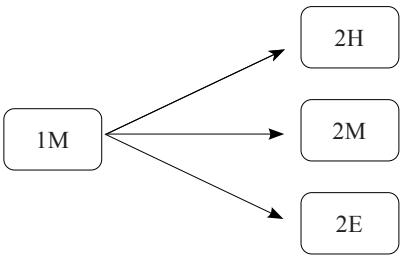
2. 多阶段自适应系统框架的搭建

获取了中文词汇测试题库中每个等级、每道题目的参数后，设计者同时对题目参数，题目所关联的背景、考核词语的词性、错误选项的分布等进行综合考量，而后对题目进行归类、组合，形成模块（Module）。由于本研究所涉个案机构的中文能力框架共八个级别，所以也相应地形成了八个级别的模块。

模块组装完毕后，一般是不同难度水平的模块混合组成阶段（Stage）。通过第一阶段数道难度中等的试题（1M）对学生进行分流，让不同中文能力水平的学生进入不同难度的第二阶段试题（2H，2M，2E）（见图1）。

3 该阅读测试为中文适性阅读能力诊断（Diagnostic Assessment of Chinese Competence，简称DACC），由国力台湾师范大学研发。

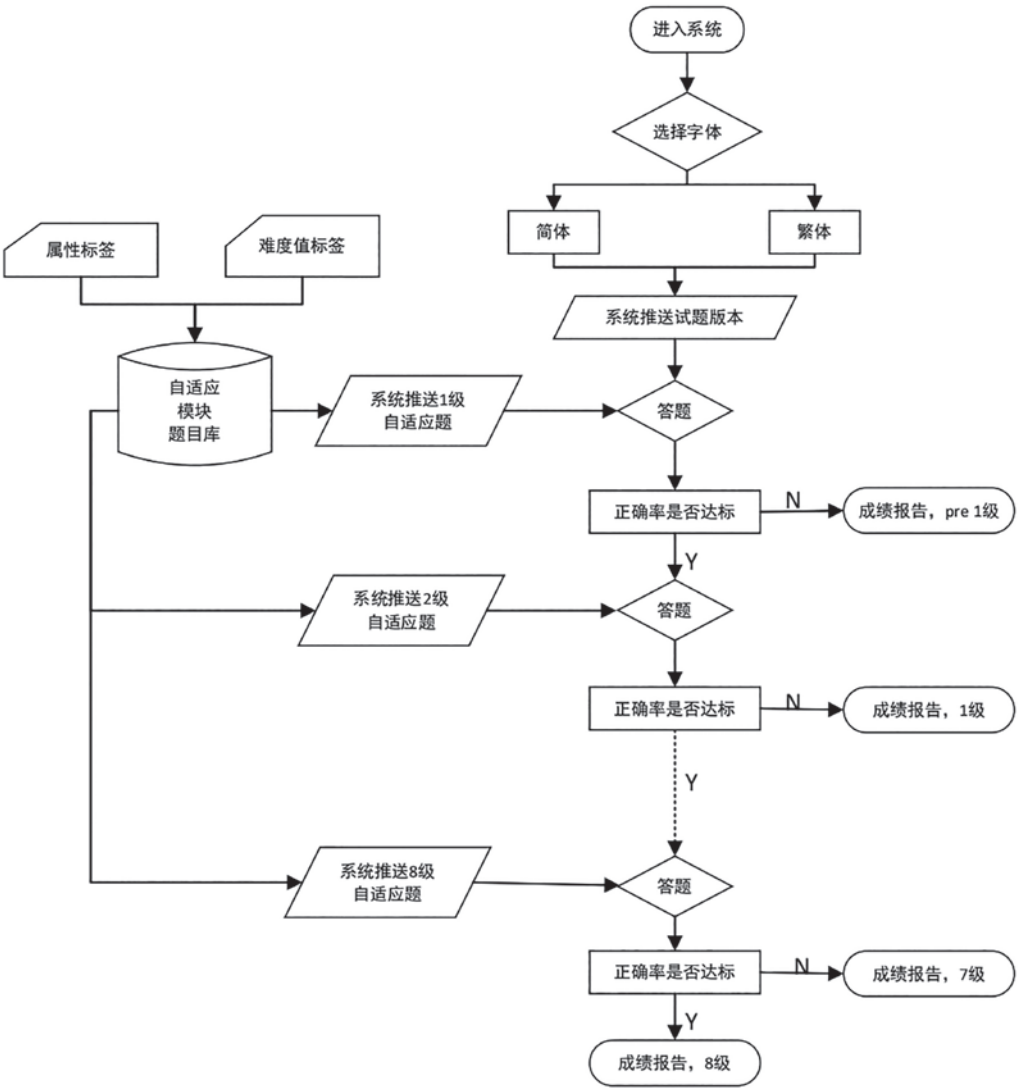
图 1：二阶段自适应测验面板



但本研究的阶段和常规 MST 定义的阶段略有不同，它并不采取以中等难度题目对学生进行分流的做法，而是所有学生，不管中文水平如何，都从最简单的一级自适应题目开始，而后进入难度逐级递升的二级、三级题组（见图 2）。所以，水平越高的学生作答的题目数越多，答题负担亦越大，反之亦然。之所以对原有多阶段自适应测试进行这样的改良，是为了给学生营造不断升级的主观感受，消除学生的考试压力。

从一个级别的自适应题目进入下一个级别的题目，需要制定路径（Path），即路由规则。路由规则是指受试者完成当前模块的作答后，进入下一个模块的规则或方法（孙小坚等，2021）。MST 中常见的路由规则有两种，一种是基于答对题目数（Number-correct, NC），一种是基于信息量（Fisher Information）。一般而言，基于答对题目数的方法较基于信息量的方法表现更好（孙小坚等，2021）。本研究即以答对题目数作为路由规则，其过程是：所有学生从一级自适应题目开始，当学生完成题目后，测试系统以 60% 作为界标，正确率低于 60% 者停留在该级，正确率大于或等于 60% 者进入二级，开始作答二级自适应题目，后续级别的考试能力分类机制如此类推（见图 2）。学生根据能力逐一作答系统推送的等级自适应题目，结束测验后，系统记录学生在所有等级的作答反应。根据学生最终停留的级别，系统将给出学生的中文词汇等级。

图 2：中文词汇多阶段自适应测试框架[4]



3. 多阶段自适应系统的部署

中文词汇能力多阶段自适应测评的框架搭建完成后，研究者将其部署在 Concerto 平台上。Concerto 是由剑桥大学心理测量中心基于 R 语言开发的开源软件，可用于开发和运行多种自适应测试。该平台在

4　图 2 出自国家知识产权局已接收的申请专利：自适应模块题目库题目质量审核和智能定级方法及系统，该专利申请方案卷号：DD22952。图 2 根据研究问题，略有改动。

数据表示、成绩评估、项目展示和外部平台资源的功能集成方面具有灵活性（Oppl et al., 2017）。利用其内置的测试流程图搭建器，研究者把不同函数的节点串在一起，可视化地编写测试逻辑，最终把所有题目部署在平台上。

4. 多阶段自适应系统的实施与对异常数据的分析

中文词汇多阶段自适应测试首次施测时间为 2022 年 9 月，学生测试时间最长约为 26 分钟。对比 2021 年 9 月使用非自适应测试系统最长耗时约 70 分钟，学生作答时间显著减少。

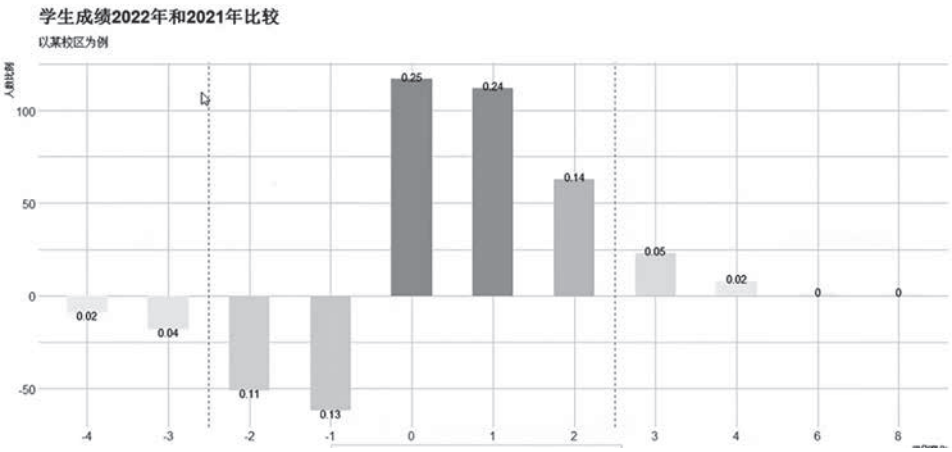
在时间缩短的情况下，本研究比较了两次测试的结果。以某校区的数据为例（见图 3），其中 0 代表 2022 年成绩与 2021 年成绩一致，1 表示 2022 年成绩比 2021 年成绩上升一级，以此类推。一般情况下，学生通过一年的中文学习理应会取得进步，这种进步可能会使他们的成绩上升到更高的级别，但也可能保留在原来的级别[5]。

2021 年采用的是非自适应系统，如出现学生成绩和等级不匹配，将请学生参加重测。如 4.1 所述，非自适应阶段是把两个等级的题目汇于一个子测试集，所以重测的考生仅能作答低 2 级或高 2 级的题目。若个别学生超出了此范畴，将无法获得更准确的判断。因此，大多数学生的成绩级别变化应该落在 “-2” 和 “+2” 之间，其中以 “0” 和 “+1” 为主。

从图三可以看出，87% 学生的等级变化落在 “-2” 和 “+2” 的范畴，13% 学生的等级变化出现异常。从统计学的角度来看，13% 的异常情况可以接受，中文词汇多阶段自适应测试结果的准确度是可信的。

5 一般而言，每个等级的学习时间约为一至两年，实际时间的长短，依据学生的个人进度而定。

图 3：某校区 2022 年和 2021 年成绩对比



多阶段自适应测试结束后，我们随机抽取了数名参与考试组织的教师，通过现场或邮件方式进行半结构访谈。访谈主要围绕学生对考试系统的体验、感受以及学生学情进行提问（见附录二），以便对测试结果中的异常数据进行成因探讨。

图 4 是进步了两个级别以上的学生，总人数为 30 人，占比 6%。结合访谈结果，分析如下：

（1）异常数据和测试方式的改变有关。图 4 中五年级以下母语生等级提高最快，分别为：三年级九名，四年级六名，五年级九名。2021 年中文词汇多阶段测试虽然是线上考试，但依然没有脱离传统的纸笔考试模态，只是从线下切换至线上。即使学生能力和等级出现落差，重测也仅限于低 2 级或高 2 级。对学生而言，能力的测评被设限了，未能映射其真正的中文能力。2022 年采取中文词汇多阶段自适应测试，考试不存在“天花板效应”，也不存在“地板效应”（杨志明、夏胜俊，2021），无论学生能力水平多高或多低，多阶段自适应题库中都具备与其能力水平相匹配的试题，以解决上述两大效应带来的困境，提高了对任何能力水平学生的判断精准度。

（2）异常数据和学生的语言背景有关。除了母语生，非母语生也同样出现等级跨度较大的情况。仍以图 4 为例，五年级、七年级和九年级均出现异常数值。结合教师访谈和学生的语言背景发现三名非母

语生分别来自韩国、日本、香港，都属于汉字文化圈的学习者，对汉字可能有着天然的优势，因此在以字词为考察对象的多阶段自适应测试里更有可能得高分，其进步甚至超过了教师的主观判断。

（3）异常数据和系统误操作有关。图 4 中出现了一名九年级学生 2022 年和 2021 年等级差距较大。结合学生个人信息及教师的访谈描述发现该生的中文能力不弱，但在 2021 年的非自适应测试中该生等级为 0，研究者推断是误操作的结果。这名学生后续可能并未参加重测，亦或是参加了重测但结果并没有更新至 2021 年的成绩单中。凡此种种，导致学生 2021 年的成绩并不具备代表性，不能反应该生真实的中文能力水平，因此 2022 年的多阶段自适应测试中等级跨度明显存在一定的合理性。

图 4：进步两个级别以上的学生

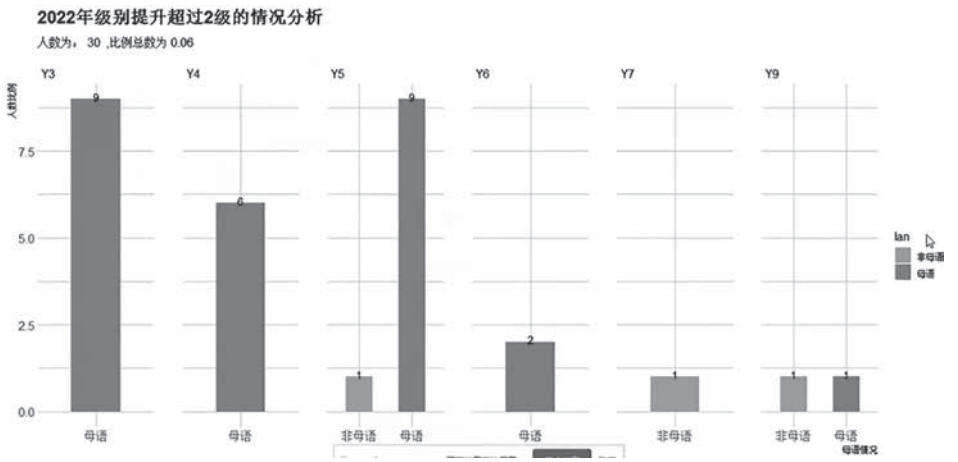
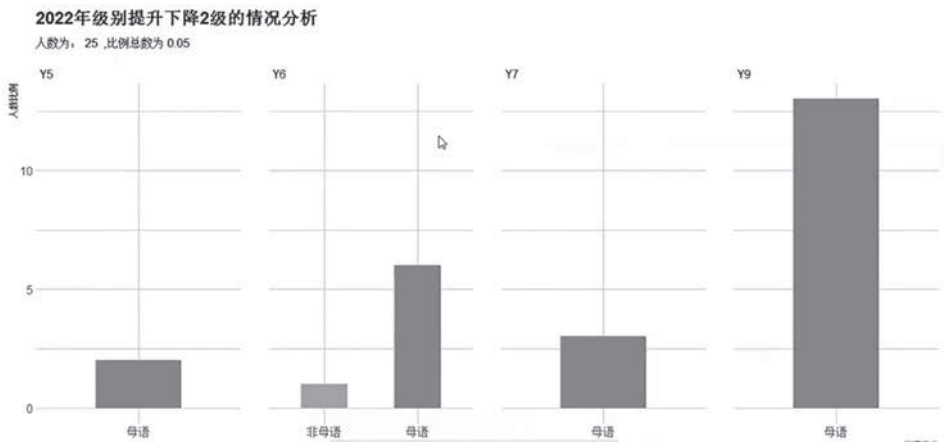


图 5 是退步了两个级别以上的学生，占比 5%，以母语生为主。从访谈结果看，这批异常数据往往和考试系统操作，学生考试态度、体验等相关。例如，部分学生想返回前一页面修改答案，于是根据日常浏览网页的习惯，直接点击浏览器“后退”按钮导致考试失效，自适应系统将其成绩定为 0 级。也有部分学生因为缺乏自适应测试的经验，在开始答题时并没有认真对待，致使考试在较低级别时便提前结束。

图 5：退步两个级别以上的学生



五、结语

本研究以中文词汇作为考试载体，参考个案机构的教学分级词表，构建了中文多阶段词汇测试题库。测试方式上，引入了多阶段自适应模式，开发了以词汇为本的多阶段自适应测试系统。结合该工具的施测结果和教师访谈，本研究发现多阶段自适应测试在用时缩短、题目减少的情况下仍能够较准确地判断学生的中文词汇能力等级。这可为教师结合其他能力维度的测试，综合判定学生中文水平提供数据，支持学生差异化学习，并进一步形成评估促进教学的良好局面。

尽管如此，在题库的构建、测试系统和考试组织上仍有亟待改进的空间，具体讨论与建议如下：

（一）题库的优化

题库的优化既有横向的拓展，又有纵向的深化。

横向拓展包括引入其他能力范畴的考核。目前多阶段自适应考试通过考察字词能力，推测学生整体的中文能力水平，这是和个案机构当下的实际需求相匹配的。但从教师教学的层面看，若想对学生的各项中文能力做更全面、细致、准确地评估，甚至是提供个性化的学习诊断信息，仍需将听说读写等能力纳入考核范畴。

纵向深化则指增加字词测试中的题目丰富性。根据个案机构四阶八级能力框架的描述，学生在第四阶段，即七级和八级，应具备“能支持自己通往更专业的各学科领域”的能力（林同飞等，2023）。为了更好地支持学生通过中文去学习其他学科的知识，并进一步学习关于中文的知识，个案机构参考“个体课文驱动”范式（金檀等，2019），编制了科学、人文、语文三门学科群的专科学术词表。相应的，题库也应考虑增加考察学科术语的题目，以此诊断学生是否具备探索学科的基本能力。

（二）测试系统的优化

“多阶段自适应”系统虽然能够依据学生的作答数据自动调整题目难度，为学生提供个性化的评估体验，但它本身的局限和不足也是显而易见的。这包括了系统稳定性和题目曝光率。

系统稳定性：多阶段自适应对算法的设计和数据的质量有较高的依赖性，算法若存在偏差，数据若不够干净都会导致结果的不稳定。

题目曝光率：系统能根据考生的能力准确推送题目，减少低难度题目对高水平考生和高难度题目对低水平考生的无效曝光。此外，算法的设定也可以在一定程度上控制题目曝光的频率，避免部分题目被过度使用。然而中等难度题目被频繁使用，低难度和高难题目曝光不足仍是多阶段自适应的现实问题。再者，频繁使用的题目容易被学生刻意记忆，也降低了考试的安全性。

因此，技术层面的优化可包括：借助首次施策的数据来校准算法，减少误判风险；引入数据增强技术，如由原来的考生手动录入信息转换为直接关联考生信息，自动登入，以提高数据质量；同时考虑题目难度和曝光率，根据两个维度对题库进行分层，轮换着使用；对曝光率特别高的题目进行降权，强制系统优选其他题目进行推送；不断增加题库内容，定期更新题目参数，降低泄题风险。

（三）考试组织的优化

无论是数据分析还是半结构访谈，均表明在中文词汇多阶段自适应测试过程中存在一些操作层面的问题，如操作不当或超时，这都会降低学生成绩的准确性。所以，在考试组织层面需要在配套的考试系统使用手册中补充误操作的说明和提醒。此外，还需要提供考试样题，让学生在进入正式测试之前对所有题型进行体验，以减少对题型理解的负担，并确保时间得到充分利用。

未来研究将在现有成果的基础上，通过优化题库、完善测试系统及改进考试组织，进一步提升多阶段自适应测试的精准性和实用性，为中文教学评估提供更科学、高效的工具支持，助力中文教育的智能化发展。

参考文献

黄山、石井秀宗（2022）：针对多元学习者的中文词汇测试题库的构建，辑于《固本求新：国际汉语教学的新理念、新思路与新方法》，（页 356-367），河内国家大学出版社。

金檀、刘康龙、吴金城（2019）：学术英语教材词表的研制范式与实践应用，《外语界》，5，21-29。

蓝珮君（2008，11 月，1-2 日）：华语文能力测验垂直等化研究，2008 台湾华语文教学年会暨研讨会，花莲慈济大学，台湾。

李贵玉、涂冬波、戴步云、宗一涛、高旭亮、苗莹（2017）：计算机多阶段自适应测验的组卷方法，《江西师范大学学报》（自然科学版），5，462-483。

林同飞、杨宝玲、陈明君（2023）：国际教育中基础教育阶段中文课程发展之个案研究—新课标的研发与分析，辑于《国际文凭课程（IB）中文教学研究新探》，（页 28-47），三联书店（香港）有限公司。

刘洪峰、郭文明、余晓佳（2016）：基于项目反应理论的自适应考试系统的研究与设计，《计算机应用与软件》，10，90-93。

刘祥友、臧志文、曾卫军等（2012）：《对外汉语偏误分析》，世界图书出版公司。

潘鸣威，孔菊芳，徐雯（2022）：高考英语（上海卷）的效标关联效度研究—来自阅读测试标准设定的证据，《外语测试与教学》，4，1-10。

孙小坚、宋乃庆、辛涛（2021）：PISA 测试中多阶段自适应测验的实施及启示，《现代教育技术》，6，72-78。

肖奚强、颜明、乔倓、周文化等（2020）：《外国留学生汉语偏误案例分析》（增订本），北京：北京大学出版社。

薛琳（2019）：《语言形态理论和英汉形态研究》，对外经济贸易大学出版社。

杨志明（2016）：计算机多阶段自适应测试探析—以《中国青少年学能发展量表》为例，《教育测量与评价》，8，4-9。

杨志明、夏胜俊（2021）："双减"背景下计算机化自适应多阶段测试的设计与算法改进，《教育测量与评价》，11，3-9。

郑蝉金、汪腾（2021）：《人工智能与智能教育丛书：计算机化自适应测验》，教育科学出版社有限公司。

Alderson,J.C. (2005). *Diagnosing Foreign Language Proficiency*. London: Continuum.

Chang,H.H., & Ying,Z.(1999). A-stratified multistage computerized adaptive testing. *Applied psychological measurement*, 23(3), 211-222.

Haladyna, T. M. , & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Ishii H, & Huang S.（2021）：Integrated Tool for Item Analysis & Response Data Analysis，检自 https://www.educa.nagoya-u.ac.jp/~ishii-h/english.html#tool，检索日期：2021.10

Kress,G.(1997). *Before Writing: rethinking the paths to literacy*. London:Routledge.

Laufer, B.(1992). How much lexis is necessary for reading comprehension? In Arnaud, P.J.L. & B é joint, H.(Eds), *Vocabulary and Applied Linguistics* (pp. 126-132). London: Macmillan.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

Meara,P., & Jones,G. (1988). Vocabulary size as a placement indicator. In P GRUNWELL (Eds.), *Applied Linguistics in Society* (pp. 80-87). London: Center for Information on Language Teaching and Research.

Nation, I.S.P.(2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Oppl, S., Reisinger, F., Eckmaier, A., & Helm, C. (2017). A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education*, 14(1),1-21.

Partchev, I., Partchev, M.I., & Suggests, M. A. S. S. (2017). Package 'irtoys'. *A collection of functions*

related to item response theory (IRT).

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Springer.

Testa, S., Toscano, A., & Rosato, R. (2018). Distractor efficiency in an item pool for a statistics classroom exam: Assessing its relation with item cognitive level classified according to Bloom's taxonomy. *Frontiers in psychology*, 9, 1585.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement*, 29(3), 243-251.

附录一：中文字词错误类型列表

偏误类型	描述 / 定义	举例
音近（phonemic）	指干扰选项和目标选项在辅音、元音或声调上比较接近。[6]	y í n —— y í ng
形近（form）	指干扰选项和目标选项在笔画或部件上比较接近。	千——干
反义（antonym）	指干扰选项和目标选项在意义（包括理性意义、色彩意义等）上相反或相对。	朋友——敌人
近义（synonym）	指干扰选项和目标选项在意义（包括理性意义、色彩意义等）上相同或相近。	坚决——坚定
同类（semantic relations）	除反义和近义之外，干扰选项和目标选项的词义还可以根据某种共同特点或关系划分出来的类。	按某种类属：桌子——椅子——板凳 按某种顺序：大学——中学——小学 按某种关系：老师——学生
多义（polysemy）	指有两个或两个以上的义项的词。	宽：①横的距离大（跟“窄”相对）；②放宽，使松缓；③不严厉，不苛求；④宽裕、宽绰。
同语素（morpheme）	指干扰选项和目标选项包含一个或多个相同的构词语素。	苍白——花白——洁白
搭配（collocation）	指词的语法功能，即该词能和哪些词组合，不能跟哪些词组合。	对某人…保密（√） 对某人…秘密（×）
最优（optimal）	指干扰选项和目标选项都可放入句子中，但从语境、意境等角度考虑，目标选项更佳。	绿竹帘子映在梳妆台镜子里，风吹着直动，（筛）进一条条阳光，满房间老虎纹，来回摇晃着。[7] 筛　　射　　透
学术（academic）	涉及比较系统且专业的学问，如法律、医生、科学等。[8]	《义务教育法》（规定）：“父母或者其他监护人必须使适龄的子女或者被监护人按时入学。”[9] 规定　　决定　　约定

6　音近、形近、同语素、搭配参考《外国留学生汉语偏误案例分析》（增订本），2020，肖奚强等，北京大学出版社。反义、近义、同类、多义参考《现代汉语》（增订四版），2007，黄伯荣、廖序东主编，高等教育出版社。

7　例子来自张爱玲中篇小说《怨女》。

8　参考《现代汉语词典》第 7 版，2016，中国社会科学院语言研究所词典编辑室编，商务印书馆。

9　例子来自《中华人民共和国义务教育法》第十一条。

附录二：访谈大纲、访谈问题

1. 关于考试现场及考试系统的体验和感受的问题：

1.1 测试系统提示语是否清晰?

1.2 需要补充或完善哪些模块?

1.3 学生常见的操作失误有哪些?

2. 有关学情分析的访谈大纲如下：

2.1 结合 GMIS 学生信息管理系统进行学情分析。

2.2 历年成绩的比对。

2.3 进展异常的学生个案分析。

2.4 总体不同年级的中文级别分布情况。

2.5 母语生不同年级的中文级别分布情况。

2.6 非母语生不同年级的中文级别分布情况。

2.7 成绩数据如何应用于教学。

The Construction and Exploration of Multi-stage Adaptive Testing for Chinese Vocabulary

CHEN, Mingjun* YANG, Baoling LAM, Tung Fei

Abstract

The case schools have transformed the relatively independent nature of first and second Chinese language curricula by developing an integrated Chinese language proficiency progression through the concept of Language and Literacy Continua. They have constructed a language proficiency framework consisting of four stages and eight levels, catering to primary and secondary school students learning Chinese regardless their language background. To align with the implementation and development of this framework, the schools have constructed a standardized testing system focusing on Chinese vocabulary. Utilizing expert judgment and item analysis methods, the study has not also built a multi-stage item bank for Chinese vocabulary but also determined the difficulty, reliability, validity and distinctiveness for each item. Additionally, to enhance test organization efficiency and improve the validity and reliability of the test, as well as to provide students with more personalized customized test, the schools have adopted a computerized multistage testing (MST) mode. This approach requires examinees with varying Chinese proficiency levels to answer item sets of different difficulties based on their responses. This paper details the construction of the test item bank, introduces the development process of the multistage adaptive testing system using item analysis, and based on testing results and teacher interviews, finds that the Chinese vocabulary multistage adaptive test can accurately estimate students' Chinese vocabulary levels while reducing both testing time and the number of items.

Keywords: Vocabulary, Multistage, Item bank, Adaptive Testing

---

* CHEN, Mingjun,Yew Chung Yew Wah Education Network.(corresponding author)
YANG, Baoling, Yew Chung Yew Wah Education Network.
LAM, Tung Fei, Yew Chung Yew Wah Education Network.