

# 国际中文计算机多阶段自适应测试的设计探索

任杰

**摘要：**国际中文测试目前以纸笔方式为主，计算机化测试为辅。虽然两种方式媒介不同，但均是一份试卷施测于所有考生。这种“多人一卷”的测试方式通常对能力分布两端的考生测量误差较大，导致整体测验信度下降。运用现代科学技术，对国际中文计算机多阶段自适应测试进行了研究设计。该测试分多个阶段进行，在跨越阶段时，系统会根据考生能力水平选择难度适当的试题模块供其作答。这种“量体裁衣”式的测试能够用更短的时间、更少的试题，更精确、有效地测量国际学生的中文语言能力。

**关键词：**国际中文；计算机多阶段自适应测试；题库

## 一、引言

国家语言文字工作委员会于 2021 年 3 月发布了《国际中文教育中文水平等级标准》。该标准的发布与实施为国际中文学习、教育教学、测试与评估提供了指导和依据<sup>[1]</sup>，有助于提升国际中文教育教学质量和效果，并进一步推动教学与测试的标准化和规范化。国际中文测试作为教学中的关键环节，能够为教学提供有效的评价和反馈，对教学具有导向性和反拨作用，进而为国际中文教育事业的发展提供强有力的支持。

当前，国际中文语言测试主要采用纸笔测试（Paper Based Test, PBT）方式，同时辅以计算机化考试（Computer Based Test, CBT）方式。在纸笔测试中，所有考生都使用同一份试卷进行测试。计算机化考试虽然使纸质试卷转化为电子形式，但仍然采用多人共用一份试卷的方式。测试的主要目的是通过最少的试题来准确评估不同考生的语言能力水平，然而，在传统的多人共用一份试卷的测试模式下，难题对于能力水平较低的考生而言往往测试效果不佳，考生在答题时容易

依赖猜测；而简单的试题则无法有效区分能力水平较高的考生。这导致测试对能力分布两端的考生存在较大的测量误差，进而影响整个测试的信效度。

随着计算机技术的发展以及项目反应理论的兴起，人们对测试的要求越来越高，“量体裁衣”式的计算机自适应测试（Computerized Adaptive Test, CAT）日益受到关注，并呈现出逐渐取代传统的多人一卷的纸笔测试的趋势。自适应性测试是指在测验实施过程中，能够根据考生的能力水平动态地调整其即将作答的试题数量和难度。在科学化题库的支撑下，CAT 首先会给考生提供一些中等难度的试题，计算机根据作答情况估计考生初始能力值，并自适应地选择一道难度适应的试题供其作答，然后再次估计考生能力值。这个过程不断循环，直至考生的能力估计值逐渐收敛，测量误差足够小，考试才会结束。CAT 的最大优点是一人一卷，每个考生的试卷难度、内容、长度均可能不同，这克服了传统纸笔测试中考生作答相同的试题而导致的某些试题无意义的

作者简介：任杰，北京语言大学语言科学院，副教授。

基金项目：国家语委“十四五”科研规划 2021 年度省部级项目“民族地区教师国家通用语言测试题库构建和计算机多阶段自适应测试研发”（YB145-12）研究成果。

现象。但在施测过程中,由于 CAT 每个试题都是线上自适应选择的,导致试题曝光、考生所考内容的不平衡性等问题突显。为此美国教育考试服务中心于 2014 年开始研发了多阶段计算机自适应测试 (Multistage Testing, MST)。区别于 CAT 的试题间的自适应, MST 以多个试题组成的模块为单位自适应抽取,同样实现了自适应的目标。由于模块可以在线下进行组装,这不仅保持了试卷内容平衡和试题的安全性,还允许考生在同一个阶段内返回修改答案。目前, MST 技术已经在 GRE、美国注册会计师测试 (AICPA)、K-12 等测验中正式使用<sup>[2]</sup>。虽然目前国内尚未见到使用 MST 技术研发并发布的测试报道,但与此相关的研究并不少。例如,关丹丹等就 PETS 进行过 MST 的研究设计<sup>[3]</sup>。近年来,北京语言大学语言学院对 MST 进行了大量研究,内容涵盖从 MST 技术的研究到系统的整体设计<sup>[4-6]</sup>。

本研究将在此基础上,进一步探索国际中文计算机多阶段自适应测试的研究设计。这种新型测试方式能够在考生不知不觉的情况下分多个阶段进行,每跨越一个阶段,系统都会根据考生的能力水平选择相应难度的试题模块供其作答。这种设计旨在以更短的时间、更少的试题数量,更精确、有效地评估国际学生的中文语言能力。

## 二、国际中文 MST 的总体规划

本研究将按照逻辑顺序,深入探讨国际中文 MST 的设计。这个过程涵盖了从预备阶段的 MST 的框架结构设计,到测试流程规划设计,再到子题库的建设和项目反应理论模型的构建方法,以及组卷方法研究。在测试实施阶段,将关注初始能力的设定、路由模块的构建、能力估计方法的选择以及路由规则的制定等关键环节。最后,在测试结束后,将对整个系统进行全面评估,并研究分数转换与报告的有效呈现方式。这一系列步骤旨在确保国际中文 MST 设计的科学性和实用性。

为便于了解整体设计,首先介绍 MST 结构设计、测试流程、所用题库等。

### (一) MST 结构设计

MST 选择 3~4 个阶段被认为是适当的<sup>[7-8]</sup>,因为这种结构可以使测验更灵活,精度更高。本研究将国际中文 MST 设计为 1-3-3 的结构,即测试分为三个阶段,每个阶段包含的模块数量不同:第一阶段仅有 1 个路由模块,第二和第三个阶段各包含难、中、易 3 个模块,具体参见图 1。这 7 个模块共同构成一个面板,该面板内共有 7 条不同路径。在第一阶段,所有考生都必须作答一个共同的路由模块,该模块由中等或混合难度的试题组成,用于估计考生的初始能力值。根据考生在第一阶段的答题情况, MST 系统会自适应地选择下一阶段中难度更高或更低的模块供考生作答。考生需要完成最后一个阶段中某个模块内的所有试题才能结束考试。为了防止试题过度曝光,可以设计组装多个平行面板。然而,增加面板数量也会相应增加所需的试题数量。本研究设计了 3 个结构相同但模块内容不同的面板,系统可以在不同面板之间灵活选择模块,从而大大减低了考生遇到完全相同的试题的可能性。最终,考生的测验总长度将是他们在三个阶段中完成的各个模块的试题数量总和,即沿着某一条路径完成的试题总数。

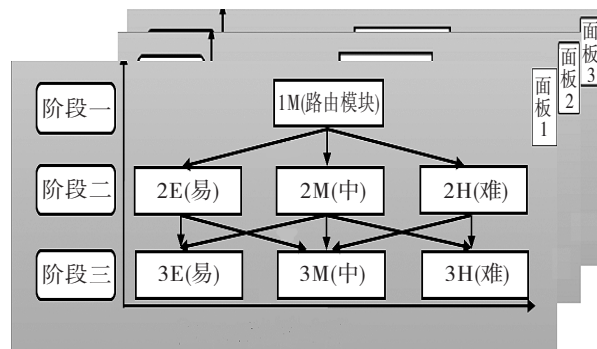


图 1 1-3-3 结构设计

### (二) 国际中文 MST 测试流程

考虑到国际中文测试包含听力分测验,需要使用听力设备,同时为了遵循语言测试习惯,本研究将听力和阅读等每个分测验视为一个独立的测验。整体测试流程如下页图 2 所示。每个测验都将进行一次 1-3-3 的多阶段自适应测试过程,先进行听力测验,再进行阅读等其他测验。

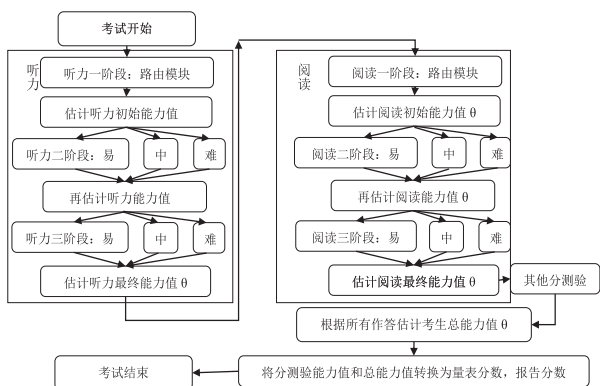


图 2 国际中文 MST 测试流程

### （三）研究所用题库

科学化的题库是 MST 得以实现的基础。本研究将采用的汉语水平测试题库（item bank）是基于项目反应理论（Item Response Theory, IRT）构建的具有统一量尺的难度、区分度等统计参数和试题的各种描述性参数以及试题构成<sup>[9]</sup>。试题部分包括题号、题组号、试卷号、分测验、题型、锚题否、锚题号、试题内容、答案、试题长度、语法点数、等级语法点数、等级词数和超纲词数等参数。描述性参数则涵盖了试题能力点、试题内容、试题功能等参数。基于 IRT 与 CTT 的统计参数包括 IRT 难度、IRT 区分度、IRT 最大信息量、IRT 最大信息点、CTT 通过率、CTT 双列和 CTT 点双列相关等。而自适应测试所用的参数则包括曝光度、回避参数以及信息总量等。

## 三、国际中文 MST 的关键环节规划设计

本研究以真实题库试题和实测试题参数为基础，分为预备阶段、测试实施阶段和测试结束后三个阶段进行各个关键环节的设计。

### （一）预备阶段

在预备阶段，除了整体测试流程和 MST 的结构设计之外，还需要深入研究子题库建设、项目反应理论模型以及模块的组卷设计。这一阶段的工作主要在线下进行，因此时间相对充裕，这有助于规避 CAT 存在的问题，从而确保 MST 的有效性和准确性。

### 1. MST 的结构设计

除了上述提及的，结构设计还需要深入研究第二、三阶段的难中易难度模块的构成、试题数量、测验时长等问题。模块内的试题数量会受到题型、可接受的信度和效度等因素的限制。为了保持一致性和标准化，同一题型的试题应该尽量被安排在同一个阶段内，并且同一阶段的模块试题数量需要保持一致。此外，测验的时长应短于纸笔测试，而具体的时长需要通过实验研究来确定。最后，为了保证不同面板间的平行，不同面板在相同位置的模块试题难度和内容等约束条件应尽量保持一致。

### 2. 项目反应理论模型的选择

项目反应理论模型有 Rasch 模型和 Logistic 模型。其中，Logistic 模型进一步细分为单参数、双参数和三参数三种类型。

在 MST 施测时，考生完成每个模块的试题测试后，将使用一种 IRT 模型来评估考生的能力。通常在研究的初始阶段，就会对纸质版试卷的实测数据进行研究，以确定最适合的模型。

### 3. 子题库建设

为了给难度不同的模块选题组卷，并降低“冷冻”题的出现概率，需要依照难度、区分度及描述性参数等关键信息，将题库进行细致的分类。在本研究中，为每个分测验精心构建了难、中、易三大子题库。同时，为了内容平衡，每个子题库还根据试题的描述性参数进一步细分为更精细的子子题库。例如，可以根据试题所测量的能力参数，将子题库进一步细分为理解和应用两个子子题库。

### 4. 组卷方法研究

由于 MST 模块和面板的组卷是在线提前完成的，因此可以很好地平衡模块测查的能力点等描述性参数，并便于控制模块难度。在组卷方法的选择上，本研究从 0-1 线性规划算法<sup>[10]</sup>和标准化加权绝对偏差启发法（NWADH）<sup>[11]</sup>中挑选了一种用于最终的模块和面板组卷。整个研究过程大致分为三个部分。首先，确定组卷的约束条件。

这涵盖了统计约束和非统计约束条件。统计约束条件主要涉及难度、区分度、信息量和试题长度；非统计约束条件则包括题型、试题考查的能力、语法、文段主题和语体等。其次，编写组卷程序。该程序从子题库中，依据组卷方法中设定的组卷规则及约束条件，自动组合成模块，并进一步将这些模块分配到每个面板的特定阶段、特定位置。在此过程中，需确保同一阶段模块的难、中、易分布均衡，且与平行面板中相对应的路径保持平行，同时还要满足每个面板每条路径的非统计约束条件。值得注意的是，不同的组卷方法会生成不同的模块和面板（测试）。最后，确定最终的组卷方法。通过模拟一组考生能力分布，将其分别施测于两种组卷方法组成的测试，然后使用各种评价方法评价两种组卷方法的优劣，最终确定本研究所用组卷方法。

## （二）测试实施阶段

在测试实施阶段，需要明确考生的初始能力设定方法，并深入研究路由模块构成、能力估计方法及路由规则等各个环节的设置。这些环节设置的合理性和准确性将直接影响测试的有效性和可靠性。

### 1. 考生初始能力值的设定

初始能力值是决定测试起点难度的关键因素，通常有三种确定初始能力值的方式。一是设定中等水平的初始值。这种方式通常是将能力值设定在中等水平，以确保测试的难度适中，既不过于简单也不过于困难。通过这种方式，测试能够更准确地评估考生的实际水平，并且为考生提供具有一定挑战性的测试内容。二是根据以往测试结果来确定初始值。这种方式是基于以往测试的数据来设定初始能力值。通过对历史数据的分析，可以了解考生的能力分布和测试难度，从而设定一个更为精确的初始值。这种方式有助于确保测试的公正性和准确性，因为它考虑到了考生的实际能力和测试的历史数据。三是考生自己选择初始值。这种方式给予考生一定的自主权，允许他们根据自己的能力和期望选择合适的初始值。这

样可以更好地满足考生的个性化需求，但同时也需要考生对自己的能力有准确的评估。考生在選擇初始值时应该谨慎，确保所选的初始值与自己的实际能力相匹配，以获得更准确的测试结果。

### 2. 路由模块

路由模块（图 1 中的 1M）在 MST 测试中扮演着至关重要的角色。作为一阶段中唯一的一个模块，它负责对所有考生进行施测，并决定考生答题路径的难易走向。由于其在整个测试过程中具有决定性的作用，关乎测验的精度，因此得名“路由模块”。一旦在路由模块中出现错误，可能会导致部分考生无法回头，这也是其得名原因之一。

对路由模块的研究涉及多个方面，其中最重要的是试题的数量、试题的难度分布以及整个模块的难度。鉴于路由模块的关键性，为确保测试的准确性和可靠性，试题数量应尽可能多。试题的难度设置通常包括中等难度和混合难度等多种组合。

### 3. 能力估计

在 MST 系统中，当考生完成一个模块的作答后，系统会立即对其能力进行估计，以便为其分配下一个难度适宜的模块。如何进行这种能力估计，是 MST 研究中一个至关重要的问题。IRT 的能力估计方法主要包括极大似然估计法（MLE）、贝叶斯期望后验估计法（EAP）以及极大后验估计法（MAP）等。

### 4. 路由规则

在 MST 系统中，每次完成考生的能力值估计后，系统会根据预先设定的路由规则自动为考生分配下一个与其能力相匹配的模块。这个路由规则设计的合理性对于确保考生的能力得到准确评估至关重要。关于路由规则的计算方法，有几种不同的选择：最大费舍尔信息法（Maximum Fisher Information）、难度匹配法（Matching b-Value）和随机选择法（Random Selection）。最大费舍尔信息法基于信息论的原理，旨在通过最大化信息量来选择最具鉴别力的试题，进而提高测试的精度。难度匹配法则主要关注试题的难度与考生能力的匹配度，通过确保试题的难度与考生的能力

水平相匹配,可以更准确地评估考生的实际水平,并有助于避免因试题难度过高或过低而导致的误差。而随机选择法相对简单,主要是随机选择试题给考生作答。虽然这种方法简单易行,但可能无法保证试题与考生能力的匹配度,从而影响测试的准确性和可靠性。

### (三) 测试结束后

测试结束后,测试系统可以获取所有面板的每条路径的流量数据和每个考生的测试路径。为了评估测试结果的优劣,需要从多个角度进行考量,包括试卷内容的覆盖度、考生能力值估计的准确性以及题库试题的使用情况等。此外,正式测试结束后,还需要为考生提供一份包含诊断信息的测验报告,以帮助考生更好地了解自己的语言水平和需要改进的地方。

#### 1. 研究结果的评价

经过一系列研究后,需要对试卷内容的覆盖度、考生能力估计的准确性和题库试题的使用情况进行全面评估,以确保测验的信度和效度。试卷内容的覆盖度是指试卷中各个知识点、能力点以及题型等的涵盖范围。一个高覆盖度的试卷能够全面评估考生的学习成果和能力水平,确保评价结果的有效性和公正性。为了衡量考生能力估计的精确程度,可以采用能力值估计的平均偏差(Bias)和均方根误差(Root Mean Squared Error,

RMSE)等评估方法,对结果进行准确的判定。由于MST模块和面板的组卷是在线下提前完成的,因此本研究对题库试题的曝光率的评估主要指的是参与组卷的试题的曝光率,即试题的使用率,以及不同考生试卷间相同试题的百分比。而题库使用情况则更多地考虑组卷后,题库中剩余试题的状况。

#### 2. 分数转换与报告

当考生完成MST所有阶段的测试后,MST会将所估计的考生能力值转化为考生能够理解的量表分数,并向考生提供包含诊断性信息的测验报告,这些报告旨在帮助考生了解自己的优势和不足之处,并为他们未来的学习和改进提供有针对性的建议。

国际中文计算机多阶段自适应测试研究是一项复杂的系统工程,需要针对每个环节进行单独的研究和设计,以确保测试的有效性和准确性。本研究所使用的汉语水平测试题库基于IRT的统计参数和完整的描述性参数,为国际中文计算机自适应测试的研发奠定了坚实的基础。通过计算机多阶段自适应测试,能够以最少的试题和最短的时间为考生提供高可信度、高效度的测试,为全球中文爱好者提供更加精准、公平的测试体验。这不仅有利于各国人民的交流和沟通,更有助于促进世界的和平与稳定。

### 参考文献:

- [1]教育部,国家语言文字工作委员会.国际中文教育中文水平等级标准[S].北京:北京语言大学出版社,2021:1.
- [2]YANDL, VON DAVIER A, LEWIS C. Computerized multistage testing theory and applications[M]. Florida, Boca Raton: CRC Press, 2014:17-18.
- [3]关丹丹,刘庆思.两种PETS计算机自适应序列测试框架比较研究[J].中国考试,2013(1):16-22.
- [4]凌好宇.计算机多阶段自适应测试若干关键技术研究[D].北京:北京语言大学,2018:27-36.
- [5]刘晴.计算机多阶段自适应测验构建中结构设计的影响因素研究[D].北京:北京语言大学,2019:26-31.
- [6]周梦梦.题库质量和考生能力分布对多阶段自适应测验精度的影响研究[D].北京:北京语言大学,2020:26-37.
- [7]LUECHT R M, NUNGESTER R J. Some practical examples of computer-adaptive sequential testing[J]. Journal of Educational Measurement, 1998(35): 229-249.
- [8]ZENISKY A, HAMBLETON R K, LUECHT R M. Multi-stage testing: issues, designs, and research[M]//VAN DER LINDEN W J, GLAS C A W. Elements of adaptive testing. New York: Springer, 2010:355-372.
- [9]任杰,黄金娜,杨翊,彭恒利.国家通用语言测验题库的构建探索[J].中国考试,2023(12):47-54.
- [10]LINDEN W J, ADEMA J J. Simultaneous assembly of multiple test forms[J]. Journal of Educational Measurement, 1998(35):185-198.
- [11]LUECHT R M. Computer-assisted test assembly using optimization heuristics[J]. Applied Psychological Measurement, 1998, 22(3): 224-236.