DOI: 10.19360/j.cnki.11-3303/g4.2006.09.002

中国汉语水平考试(HSK)和中国少数民族汉语水平等级考试(MHK)都是为测试母语非汉语者的汉语水平而设立的国家级标准化考试。这两项考试是由北京语言大学汉语水平考试中心设计研制考试,正在迅速发展,发挥着越来越重要的作用。北京语言大学汉语水平考试中心在控制这两项考试质量的工作流程中,引入了《必查项目表》,完善了审题制度,改进了试卷等值方式,设定了合理的及格分数线。

HSK 和 MHK 在考试质量方面的探索

□ 北京语言大学 谢小庆

“中国汉语水平考试(HSK)”和“中国少数民族汉语水平等级考试(MHK)”都是为测试母语非汉语者的汉语水平而设立的国家级标准化考试。HSK 目前在世界的 38 个国家设立了 100 多个考点,在国内 31 个城市设立了 56 个考点。伴随中国经济规模的迅速扩大和国际影响力的增加,HSK 正在迅速发展。今天,MHK(三级)已经成为新疆、吉林、青海、四川等省区的高考科目,在促进少数民族汉语教学和推动西部教育发展方面,在推动少数民族地区以素质教育为取向的课程改革方面,发挥着越来越重要的作用。

考试是一把“尺子”,或一个“体温计”,被用来测量人的能力或其他心理特征。考试这把尺子并不是天然可靠的,可能存在误差,有时甚至是较大的误差。考试的质量控制过程就是控制考试误差的过程,就是对考试的效度、信度、公平性、分数等值性、安全性进行质量检验的过程。在质量控制方面,我们主要参考的检验标准是美国教育研究协会、美国心理协会和美国国家教育测量协会三家共同颁布的《教育与心理测量标准》。在这一标准中,包含对考试的效度、信度、等值、分数报告、档案管理、考生权益的保护、考试使

用者应尽的义务等方面较详细的指导性意见。

HSK 和 MHK 质量控制的主要环节包括：

● 测验标准的制定是否基于工作或活动分析？

● 测验内容的确定是否符合测验标准？

● 测验的形式、题型是否经过专家充分的论证？

● 命题人员是否具备资格？

● 命制的试题是否经过必要审查？

● 正式考试前是否经过预测？

● 预测样本是否具有代表性？

● 对试题及试卷的统计分析指标是否达到要求？

● 测验长度的确定是否具有合理依据？

● 及格线的确定是否具有合理依据？是否经过充分论证？

● 试卷的印刷、运输等各个环节是否安全？

● 对主、监考人员是否进行了充分的培训？

● 考场纪律是否得到维护？是否有效地防止了作弊现象？

● 考生的答题情况是否得到准确的录入？

● 阿尔法系数、分半信度等信度指标是否达到要求？

● 是否得到充分的效度证据支持？

● 对不同性别、民族、地域、经济条件的考生是否公平？是否存在 DIF 和 DTF？

● 不同版本测试的分数之间是否具有可比性？是否经过等值处理？

● 测验分数是否得到了准确、充分的解释？测验分数是否具有合理的参照系？（常模参照和标准参照）

● 是否存在对分数的误解？是否存在分数的误用现象？

● 是否建立了足够规模的题库？

从 1984 年开始研制以来，HSK 在提高考试质量方面已经摸索了 20 年。以往，我们在编制作为考试依据的汉字和词汇等级大纲，借助词频统计确定考试内容，注重考查汉语交际能力，坚持考前预测，坚持统计等值，控制作文口试的主观评分误差，题目功能差异(DIF)分析等方面，进行了一些探索。这些探索，对于提高 HSK 和 MHK 的质量产生了重要的作用。

在 2005 年，我们在提高考试质量方面又做出了一些新的努力。

首先，在工作流程中引入《必查项目表》。为了保证考试的质量，我们制定了明确的操作规范。每个考试在命题过程中都有一份《必查项目表》，其中的内容包括卷号是否正确，题号是否正确，各部分是否齐全，题数是否正确，选项标号是否正确，等等。在听力录音带的《必查项目表》中，包括题号是否正确，说明是否齐全，等等。命题人对每份试卷都要逐项检查，检查之后逐项签名，研究室主任也要逐项签名。严格按照操作规范进行操作，对于我们保证试卷质量是很重要的。

其次，完善审题制度。我们完善了较严格的审题制度。HSK 是三审制度，完成《必查项目表》是一审，是汉语水平考试中心自己进行的审题。之后进行二审。我们聘请北师大、人大等学校的专家参加二审。二审之后进行预测。在对预测结果进行统计分析的基础上，进行拼卷。拼成的试卷由国家汉办组织三审，即终审。

MHK 所执行的是四审制度。四审主要是敏感性审查，检查是否存在冒犯少数民族习俗的问题。经过预测以后拼成的正式试卷要送到民族大学进行四审，请 4 位教授参加，一位蒙古族、一位朝鲜族、一位藏族、一位维吾尔族。四审仅仅审查冒犯民族习俗方面的问题，不涉及其他语言、内容等方面的问题。审查的原则是“一票否决”，不讨论。只要有教授提出异议，就将题目删除。

第三，改进试卷等值方式。我们最初的等值模型是 Tucker 线性模型。后来，我们开始引入了 Levine 观察分数线性模型、Levine 真分数线性模型和等百分位模型等不同的等值模型，互相参照。1998 年，我们初步建成了基于现代项目反应理论（IRT）的题库。借助计算机从题库中半自动化生成的试卷被应用于正式考试，实现了试卷的 IRT 等值。与此同时，我们也开始在小样本试卷等值中酌情采用平均数等值。近年，为了克服共同题曝光带来的负面影响，我们开始引入共同组设计的线性等值，对多份试卷进行了不含共同题的共同组线性等值。为了保证考生的动机水平，我们在报名时就明确宣布：考生可以自愿参加两次考试，其中一次免费。我们将报告两次成绩中较好的一次。两次考试安排在上午和下午。每一组 500 人。在一次考试中我们可以安排多组人参加上、下午的两次考试。有时候，在同一天的一次考试中我们就可以完成多份新试卷的等值数据收集工作。

我们还引入了分半组卷的混合等值设计。具体作法是：将“标准卷”以分测验为单位分成两部分。同时，将新卷也以分测验为单位分成相应的两个部分。将标准卷和新卷交叉组成两份用于等值的试卷。对于新试卷来说，这是一个经过处理的共同组等值。对于参加两份等值测试的考生来说，这是一个共同题等值，可以在经过等值处理后，即刻报告成绩。这两个试卷可以在正式考试中应用。在同一次大规模考试中，两个“分半组合”的等值试卷可以同完全不包含共同题的新卷同时施测。考试之后，可以立即通过两个等值卷计算出新卷与标准卷的等值关系，及时报告成绩。

第四，设定合理的及格分数线。HSK（初中等）和 MHK（三级）的一个重要用途是考查一个母语非汉语的大学生是否可以开始汉语授课的专业学习。及格线的划定不应是某一预先设定的通过比例，而应基于对专业学习活动的分析之上。尽管教育部文件已经规定了 HSK 三级是进入理工农医专业学习的标准，HSK 六级是进入文史哲中医专业学习的标准，但由于这一标准的制定缺乏科学依据，许多高等院校并没有在实践中执行。我们采用 Angoff 方法等一些更合理的方法对 HSK 的及格线进行了验证性研究。在 MHK 的及格线建立过程中，也采用了包括 Angoff 方法在内的一些更合理的方法。

第五，对测验分数做出“能做”解释。任何一组考试题目或考试任务，都可以得到一个“正确回答数”或“正确回答比例”，都可以得到一个“分数”。但是，并不是任何一个分数都可以根据考试的目的做出解释，都可以被赋予“意义”，都可以成为决策的合理依据。对于 HSK 和 MHK 分数的使用者（主要是招生、招工者）来说，所关心的并不是分数本身，而是分数中所包含的关于考生“能做什么”的信息。缺乏必要的关于“能做什么”的分数解释，是以往 HSK 分数报告中的明显不足之处。对此，我们已经开始借鉴 ETS 的一些经验，对获得不同 HSK 和 MHK 成绩的考生进行活动分析和汉语实际应用能力调查，开始尝试对 HSK 和 MHK 分数做出更丰富的“能做”解释。

今年，我们继续对 HSK 和 MHK 的质量进行完善，包括：借鉴欧盟、美国、加拿大和澳大利亚等国的经验，建立明确的汉语作为第二语言的评价标准；探索基于语料库的考试命题方式，建设和完善汉语语料库、汉语中介语语料库、作文语料库和口语语料库；完善题库建设，在计算机自动组卷过程中引入更多的内容控制变量；引入“预先等值设计”，在正式考试中包含预测题目，将预测和等值在正式考试中同时实现；加速基于计算机和基于网络的 HSK 和 MHK 研究，加速自适应性考试研究，争取早日付诸使用；加速“作文电子评分员”研究，争取早日付诸使用等。