

谢小庆 彭恒利 / 主编

考试研究文集

(第2辑)

科学出版社

考试研究文集

(第2辑)

谢小庆 彭恒利 主编

经济科学出版社

图书在版编目 (CIP) 数据

考试研究文集. 第2辑/谢小庆主编. —北京: 经济科学出版社, 2004.4

ISBN 7-5058-4026-6

I. 考… II. 谢… III. 汉语-水平考试-研究-文集 IV. H195-53

中国版本图书馆 CIP 数据核字 (2004) 第 019977 号

责任编辑：卢元孝
责任校对：徐领弟
版式设计：代小卫
技术编辑：潘泽新

考试研究文集（第2辑）

谢小庆 彭恒利 主编

经济科学出版社出版、发行 新华书店经销

社址：北京海淀区阜成路甲28号 邮编：100036

总编室电话：88191217 发行部电话：88191540

网址：www.esp.com.cn

电子邮件：esp@esp.com.cn

天宇星印刷厂印刷

德利装订厂装订

880×1230 32开 12.5印张 330000字

2004年4月第一版 2004年4月第一次印刷

ISBN 7-5058-4026-6/F·3184 定价：20.00元

（图书出现印装问题，本社负责调换）

（版权所有 翻印必究）

前 言

2003年4月，一场不期而遇的灾难降临到北京，一场没有硝烟的战斗在北京打响。许多单位中止了正常的工作。曾经拥堵的北京街头，顿失喧嚣。曾经热闹的北京语言大学校园，倏然安静。但是，在这段时间，在安静冷清、人影稀少的办公楼四层，在汉语水平考试中心MHK研究室中，却几乎每天都是紧张工作的场面。MHK课题组的研究人员和工作人员，像平时一样地每天坚持上班。一方面，为即将如期在6月初举行的新疆高考汉语试卷进行命题、审题、组卷、录音等工作，一方面，按照工作计划对已经完成预测的MHK多套预测试卷进行试题分析，拼组正式试卷。

在这个“非典型”的时期中，更凸显了MHK课题组成员们的敬业精神。这种敬业精神，也体现在这本汇集课题组科学研究成果的文集中。

MHK课题组成员的基本信念是：考试必须建立在科学研究的基础之上，必须通过科学研究来保证考试的质量，考试命题不能靠“眉头一皱、计上心来”。所有参加考试的考生都应该受到完全相同的对待，他们的成绩应仅仅反映他们的汉语水平，而不应该受到试卷质量、试卷难

度的影响。MHK 的证书授予应该具有一致的、稳定的标准,这种标准不应受到试卷本身质量的影响。虽然我们很难保证每一个达到规定标准的考生都能够获得 MHK 证书,还很难保证每一个未达到规定标准的考生都不能获得 MHK 证书,但这是我们不倦追求的目标。我们努力将误差的范围压缩到最小的程度。

基于这些信念,我们进行了一系列的研究和思考,内容覆盖能力与知识的关系,能力考试与知识考试的关系,对考试公平性的理解,对测验效度的理解,对什么是语言能力的理解,对什么是标准化考试的理解,编写“汉语语法大纲”的必要性,怎样编写汉字和词汇大纲,怎样更准确、有效地考查听力、阅读、书面表达和口语能力,怎样做好命题和审题工作,怎样实现不同难度试卷之间的等值,怎样合理地确立及格分数线,怎样合理地确定测验考试的长度,怎样赋予测验分数更丰富的解释,等等。

我们知道,不仅仅 MHK 课题组具有这些关于考试科学性和公正性的信念,许许多多考试工作者们都与我们具有同样的信念。我们将这些思考和探索编辑成册,是希望我们的这些努力,能够对考试同行们有所启发。我们也希望借此与为数众多的考试工作者们展开更广泛的交流,听到同行们对我们的指正和教诲。

编者

2003 年 10 月 14 日

目 录

● 概 论 ●

对测验效度的一些新认识.....	谢小庆 (1)
关于考试公平性的一些思考.....	谢小庆 王 洋 (9)
HSK 的教育测量学研究.....	谢小庆 (18)
能力与能力考试.....	谢小庆 (31)
谈语言能力的考查.....	谢小庆 彭恒利 (38)
标准化测验在理解上的一致与分歧.....	徐 静 王瑞峰 (51)
关于制订“汉语语法大纲”的建议.....	程 棠 (60)
刍议 MHK 词汇大纲.....	张晋军 任春燕 (70)
语文词典词性标注的几个问题.....	姜德梧 (79)
汉语测试电子评分员研究设想.....	张晋军 任 杰 钱 雪 (90)
中美考试质量控制机制的比较.....	谢小庆 (113)
关于听力理解能力不同层次的思考.....	鲁新民 (121)
民族汉考 (MHK) 的命题和审题.....	彭恒利 张晋军 (134)
民族汉考 (三级) 命题注意事项.....	张晋军 彭恒利 (147)

● 实证研究 ●

新疆 2002 年、2003 年高考汉语科目考试改革

述评 鲁新民 彭恒利 (155)

2002 年、2003 年新疆、内蒙古、延边三地 MHK

预测卷作文字数统计的一些分析 张晋军 任春燕 (169)

关于统计等值效果的系列试验研究 谢小庆 (176)

民族汉考三级分界标准确立初探 武晓宇 徐 静 赵 玥 (187)

关于民族汉考测验长度的实证研究 柴省三 (199)

汉语水平考试 J324 卷构想效度的验证研究 李 慧 朱军梅 (209)

HSK 作文评分客观化探讨 任春艳 (218)

关于“新闻听力检测书面问答与声音问答效果

比较研究”的实验报告 刘士勤 (237)

对朝鲜族被试汉语篇章衔接方式使用情况

的考察和分析 彭恒利 (248)

对 HSK (初中等) 预测结果真实性的验证性研究 许雪立 (271)

HSK 在海外: 状况、问题和对策 许义强 (281)

● 文献综述 ●

国外计算机自适应性测验 (CAT) 选题策略研究

的新进展 张忠华 谢小庆 郑日昌 (299)

美国教育考试与评价现状 齐 森 (315)

“托业”的分数解释体系及其启示 王 洋 (330)

分界标准研究的新进展 赵 玥 (344)

中国语鉴定考试介绍 李桂梅 (361)

❁ 其他 ❁

公务员需要什么样的语言能力 张晋军 (373)

❁ 附录 ❁

教育部关于在有关省区试行中国少数民族
汉语水平等级考试的通知 (387)

对测验效度的一些新认识

谢小庆

摘要：本文介绍了美国《教育与心理测试标准》1999 年新版中关于效度的定义和分类。在这一文献中，效度被定义为测验对构念所测量的程度。讨论了从内容、反应方式、与其他变量之间的关系、内部结构、测验结果等多个渠道积累效度证据的重要意义。

关键词：测验 效度

考试是一把尺子，被用来测量考生的能力。这把尺子本身可能存在质量问题。只有达到质量标准的考试才能被应用。坚持这把“尺子”的质量标准，不仅是为了通过考试把最优秀的人才选拔进学校、机关，而且是为了维护社会公正。根据一把质量不高的尺子将一些人拒之于学校、机构的大门之外，是不公正的。

效度，即有效性，是刻画考试质量的最重要指标之一，它反映了考试在多大程度上实现了考试目的。人们对于效度问题的认识是

逐渐发展和深化的。国际心理测量学界对于效度问题的一些新认识体现在 1999 年修订的新版《教育与心理测量标准》（standards for educational and psychological testing，以下简称《标准》）中。《标准》是由美国教育研究协会（AERA）、美国心理学会（APA）和美国国家教育测量协会（NCME）三家联合颁布的。这一版本是对 1985 年版本的修订，从 1985 年的 100 页增加到 1999 年的 194 页，增加了许多内容。

一、效度概念的定义

在《标准》1985 年版本中，效度被定义为“从测验所做出推论的适当性或合理性的程度”（第 94 页）。“效度反映已有证据可以在多大程度上支持根据测验分数所做出的推论。”（第 9 页）根据证据来源不同，证据被划分为来自“构念（construct）”、来自内容和来自标准（criterion）三种，效度也被相应地划分为三种。多年来，这种关于效度的定义和效度种类的划分，一直成为教育与心理测量学界关于效度研究的基本框架。

在《标准》1999 年新版本中，在效度定义和效度种类划分方面，出现了明显的变化。新版本中，效度被定义为“关于测验分数的特定解释所得到的支持程度。这种支持来自累积的证据或理论。这种解释是测验应用的基础。”（第 184 页）

“逻辑上，效度估计始于对测验分数如何解释的清晰说明，以及一个关于分数解释与测验应用之间关系的说明。所谓测验解释，是关于测验所要测量的构念（construct）或概念（concepts）的解释”。（第 9 页）“在本标准中，所有的分数都被视为对构念的测量。”（第 174 页）

编制一个测验，首先需要回答的问题就是：“这个测验测什么？”对这个问题的回答，就是“构念”。例如，一个汉语水平考试

测量的是“汉语能力”，一个数学能力测验测的是“数学能力”，一个焦虑性测验测的是“焦虑”。这里，“语言能力”、“数学能力”、“焦虑”等就是“构念”，就是研究者为了对这些问题进行研究而构造出来的一些概念。

从 1955 年 Cronbach 与 Meehl 提出 construct validity 概念以后，心理测量学家对这一概念就存在两种不同的看法。反对的人认为它会导致对测验效度的主观臆测，支持的人认为它涵盖了所有其他的效度证据。从《标准》1999 年版本看，后一种观点今天已经占据了明显的上风，construct 已经成为教育与心理测量中最重要、最核心的概念之一。这里，构念将不再是效度证据三种来源之中的一种，而是被用来定义效度概念。这一改变表明，在主流教育与心理测量学界，今后已经不再存在“构念效度 (construct validity)”这一概念。所谓效度，就是测验对构念进行测量的有效程度。因此，“构念效度”“这一短语对于效度来讲已经成为多余 (redundant)” (第 174 页)。随着“构念效度”这一概念退出历史舞台，“构念”概念却走到了舞台的中心。

在 1985 年版《标准》中，construct 被定义为“不可直接观察的、体现为个别差异的心理特征”。在 1999 年新版中，construct 被定义为“测验所要测量的概念或特性 (the concept or the characteristic that a test is designed to measure)” (第 173 页)。在中文文献中，目前对于 construct 的中文译法主要有“结构” (朱智贤)、“构想” (朱智贤、张敏强、郑日昌等)、“概念” (谢小庆, 1988)、“实验” (桂诗春)、“构造概念” (凌文轮等)、“建构” (简茂发)、“构念” (杨宜音等)、“构卷” (高兰生等) 等。综合权衡，笔者认为“构念”译法较好，(谢小庆, 2001) 本文采用了这种译法。为了与 standard (标准) 相区分，本文将 criterion 译为“效度标准”。

二、效度证据的来源

根据 1985 年版本，效度证据来源于构念、内容和效度标准三个方面。在新版《标准》中，没有再沿用这种关于效度的分类，而是讨论了多种效度证据的来源。

（一）基于内容的证据（evidence based on content）

通过考察测验内容与测验构念之间的关系，可以得到重要的效度证据。通常，在测验编制之前都需要对测验的内容范围进行界定、分类，并确定各部分内容的比例。通过系统比较一份测验的实际内容与测验说明中对测验内容的界定，通过比较实际测验各个部分的内容比例与测验说明所确定的比例，可以得到重要的效度证据。内容效度通常以专家评定的方式进行，请专家对测验的各个部分与测验构念之间的关系进行系统评价。专家可以对测验题目的覆盖程度进行评价，也可以对各部分内容的相对比重或相对重要性进行评价。评价内容效度可以体现为测验所包含内容对一个内容总体的代表性，如中学物理考试；也可以体现为测验所包含的一组任务对一个任务总体的代表性，如美容师资格考试。

（二）基于反应过程的证据（evidence based on response processes）

通过对测验参加者的反应方式进行分析，可以得到关于测验效度的证据。这种反应模式分析可以对测验“所测”与测验“欲测”之间的一致性提供支持。例如，一个测验的“欲测”构念是“推理能力”。一份实际的测验，可能确实考查的是考生的推理能力，具有较好的效度；也可能仅仅反映出考生对一些特定知识或结论的记忆程度，效度不高。又如，一个测验的“欲测”构念是“内外向”。

一份有效的测验，受测者的得分可以反映出受测者的“内外向”；一份缺乏效度的测验，受测者得分可能很大程度上受到社会称许性的影响。通过对考生的解题过程和解题策略的调查，可以得到有关考生反应过程的信息。因此，可以通过要求考生记录解题过程、描述解题策略等方式进行效度研究。在有的测验中，来自反应时间和眼动仪的资料可以成为效度证据。

在一些包含主观性评分的测验中，反应过程就不仅仅是考生的反应，还会受到评分人的影响。在这类测验中，测验的分数就受到考生和评分人两个方面的影响。这时，对评分人评分过程的考察可以为测验提供效度证据。例如，在一个作文评分标准中包含了“感情真挚”一项。如果作文考试所考查的构念是“语言能力”，那么，感情是否真挚就似乎与测验目标关系不大。在这类情况下，通过考察评分者在评分过程中所考虑的因素，可以为考试的效度提供支持。

(三) 基于内部结构的资料 (evidence based on internal structure)

许多测验都是基于某种特定的理论框架之上的。有的测验是单维的，如《瑞文推理测验》。有的测验是多维的，如《16种人格因素问卷》。不论测验是单维或多维，通过对测验内部结构的考查，都可以得到关于测验效度的证据。在测验是单维的时候，测验的同质性指标（如 α 系数、因素分析的第一主因素所解释的方差比例等）就既是测验信度的指标，也是测验效度的指标。在测验是多维的时候，通过相关分析、因素分析等方法可以得到关于测验内部结构的信息，这种信息显然是关于测验效度的重要证据。题目功能差异（differential item functioning, DIF）方面的研究也属于测验内部结构方面的研究，也可以对测验的效度提供支持。

（四）基于与其他变量之间关系的资料（evidence based on relations to other variables）

考察测验分数与独立于测验的其他变量之间的关系，可以得到最重要的效度证据。效度证据首先可以来自于反映测验目的的效度标准。这时的基本问题是：测验可以在多大程度上预测效度标准或反映效度标准？当测验先于效度标准发生的时候，我们关注预测效度。例如，在对大学入学考试的效度进行研究的时候，常常将考生入学后的在校成绩作为效度标准，计算入学成绩与在校成绩之间的相关系数，以此来估计入学成绩可以在多大程度上预测在校成绩。当测验数据与效度标准数据同时取得的时候，我们关注同时性效度。例如，我们在检验一个以选择题方式进行的文字表达测验的效度时，可以同时施测一项主观评分的作文考试。以后者作为效度标准，检验前者的效度。

效度证据也可以来自于测验与其他一些测验的关系。这些测验可以是与本测验测量相似构念的测验。例如，一个新编的成人智力测验与权威性智力测验《韦氏成人智力量表》之间得分的一致性，可以是对新测验的一种效度支持。这样的效度证据被称为会聚性证据（convergent evidence）。也可以是一些与本测验测量不同构念的测验。例如，如果一个“言语测验”与“推理测验”之间具有很低的相关，这一结果是对这个言语测验效度的支持，说明这个测验确实是一个“言语测验”而不是一个“推理测验”。关于一个新测验“不测什么”的信息，也是一种效度证据，被称为区分性证据（discriminant evidence）。

区分性效度证据可以来自于精心设计的实验研究。例如，效度证据可以来自于对比辅导研究。对于一个用于入学招生的能力测验，如果短期强化辅导可以明显提高学生成绩，测验效度就应该受到质疑。

(五) 基于测验结果的资料 (evidence based on consequences of testing)

关于测验结果的分析可以为测验效度提供支持。测验的使用会带来种种结果,有些是期望之中的,有些是与期望无关的。例如,在职工招聘中使用测验可以导致培训费用的降低和工作效率的提高;在学校招生中使用测验不仅有助于保证教学秩序,而且具有极大的激励作用;激烈的考试竞争造成学生学习负担过重以致影响到学生个性的健康发展;一些家庭经济条件较差的学生可能受到不公平对待,等等。造成这些结果的原因是复杂的。许多结果并不是由测验造成的。尽管如此,对测验结果的分析常常可以为考试的效度提供有利或不利的证据。在这种结果分析中,需要认真区分哪些结果是由测验造成的,哪些结果并不是由测验造成的。例如,激烈竞争的升学考试,可能导致一些学生死记硬背知识性结论而忽视发展解决问题的能力。这类后果,可能与考试的效度有关,需要通过改进考试效度来解决。激烈竞争的考试还可能导致学生个性发展方面的问题。这类后果,可能与考试的效度关系并不大,需要通过改进整个招生考试制度来解决。

综上所述,从1999年新版《标准》中可以看出,效度被重新定义为测验对构念所测量的程度。在效度的分类方面,已经放弃了1985年版《标准》中构念效度、内容效度、效度标准关联效度的划分,而是从内容、反应方式、与其他变量的关系、内部结构、测验结果等几个方面讨论了测验效度证据的来源。在新版《标准》中特别强调了从多种渠道积累效度证据的重要性。通过效度证据的不断积累,我们将更恰当地使用测验分数,更准确地对测验分数进行解释,将对测验构念的定义不断完善,将对测验本身不断地进行修订和完善。同时,在效度证据积累的过程中,我们可以发现和提出新的需要研究的问题。新版《标准》特别指出,测验的效度依赖于测验的精心编制,依赖于测验编制的理论框架,依赖于测验的施测和计分过程,依赖于分数等值,

依赖于及时纠正测验过程中出现的不公平因素, 等等。

参考文献

- [1] American Educational Research Association (1985) Standards for educational and psychological testing
- [2] American Educational Research Association (1999) Standards for educational and psychological testing
- [3] 高兰生、陈辉岳 (1996):《英语测试论》, 广西教育出版社, 第 74 页
- [4] 桂诗春 (1986):《标准化考试——理论、原则与方法》, 广东高等教育出版社, 第 140 页
- [5] 简茂发 (1997):《心理与测验统计方法》, 台湾, 心理出版社, 第 25 页
- [6] 凌文轮、滨治世 (1988):《心理测验法》, 科学出版社, 第 16 页
- [7] 谢小庆 (1988):《心理测量学讲义》, 华中师范大学出版社, 第 173 页
- [8] 谢小庆 (2001):《关于 construct 的译法》,《心理学探新》, 21-1, 第 64 页
- [9] 杨宜音等译 (1998):《性格与社会心理测量总览》, 台湾, 远流出版公司, 第 965 页
- [10] 张敏强 (1998):《教育测量学》, 人民教育出版社, 第 123 页
- [11] 郑日昌等 (1990):《考试的教育测量学基础》, 高等教育出版社, 第 83 页
- [12] 朱智贤 (1989):《心理学大辞典》, 北京师范大学出版社, 第 240、331 页

Insight of Test Validity

Abstract: The revised definition in the 1999 edition of Standards for Educational and Psychological Testing gives us a better insight of the fundamental concept, which in turn has many important implications for the design of tests and the interpretation and use of test score.

Key words: testing, validity