考试研究文集

（第4辑）

谢小庆　张　洁／主编

经济科学出版社

考试研究文集

（第4辑）

谢小庆　张　洁　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4辑/ 谢小庆，张洁主编．—北京：经济科学出版社，2008.5
ISBN 978-7-5058-7138-0

Ⅰ．考… Ⅱ．①谢…②张… Ⅲ．考试学-文集 Ⅳ．G424.74-53

中国版本图书馆CIP数据核字（2008）第055543号

责任编辑：唐俊南　卢元孝
责任校对：徐领柱
版式设计：代小卫
技术编辑：潘泽新

考试研究文集（第4辑）
谢小庆　张　洁　主编
经济科学出版社出版、发行　新华书店经销
社址：北京市海淀区阜成路甲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
汉德鼎印刷厂印刷
永胜装订厂装订
880×1230　32开　15印张　400000字
2008年6月第1版　2008年6月第1次印刷
ISBN 978-7-5058-7138-0/F·6389　定价：35.00元
（图书出现印装问题，本社负责调换）

前　言

根据新华社2008年2月3日报道，过去一年中珠江三角洲制鞋企业关闭千余家（参看2008年2月4日《南方日报》A13版）。这一现象不仅折射出珠三角地区加工贸易正在经历着转型的阵痛，也折射出中国经济发展所面临的挑战和机遇。

30年来，中国确实创造了经济发展的“奇迹”。但是，这种主要凭借劳动密集、廉价劳动力、低人权标准、高污染容忍等方面“优势”带来的经济发展，并没有能够在高科技领域中缩小我国与发达国家的距离，相反，距离还在继续加大。苏联和中国分别在1949年和1964年爆炸了自己的第一颗原子弹。1961年，苏联实现了首次载人航天飞行。2003年“神舟5号”的发射，意味着中国与俄罗斯的差距从1964年的15年扩大到2003年的42年。1958年，美国发射了自己的第一颗人造地球卫星。1970年，中国发射了自己的第一颗人造地球卫星。1968年，美国宇航员阿姆斯特朗踏上了月球。据我所知，到2008年为止，尽管中国的“奔月”计划进展顺利，但中国的宇航员尚未踏上月球。这意味着中国与美国的差距已经从1970年的12年扩大到2008年的至少40年。

影响我国成功实现经济转型和产业升级的一个重要因素是教育。产业升级的最重要因素是人才，而人才主要靠教育来培养。如果不能尽快实现“应试教育”向“素质教育”的转变，如果不能尽快实现儿童从“厌学”向“爱学”的转变，如果不能尽快改变“教一本书、学一本书、背一本书、考一本书”的教育现状，中国就可能错失经济转型和产业升级的机遇。

之所以存在“应试教育”与“素质教育”的矛盾，原因就在于考试所考查的是“记忆”而不是“素质”。一旦考试所考查的是“素质”，“应试教育”与“素质教育”就实现了统一。

我们坚持认为，对于21世纪中国的发展，今天没有一件事情比2001年开始的基础教育课程改革更重要。她的成败，关系到今后几十年中国的命运。然而，她的成败很大程度要取决于考试评价制度的改革。如果没有考试评价制度的改革，“新课程改革”终究会前功尽弃，所有的努力都会回到原点。7年来，我们看到许许多多中小学老师们为了挽救小“范进”们而辛辛苦苦地推动着“新课程改革”，我们对他们怀着深深的敬意。但我们清楚地知道，如果评价制度不改，如果高考不改，他们的所有努力都会付诸东流。

每每想到这些，我们就感到自己作为专业考试研究者的责任，我们就产生大声呐喊的冲动。像研究文集前3辑一样，在本辑中记录了我们在探索考试科学化方面的一些脚步，凝结了我们为了推进考试的科学化改革所进行的努力，发出了我们近乎声嘶力竭的呼喊。

从本辑中可以反映出我们在考试科学化方面所进行的一些新探索。例如，在题库中试题内容参数体系构造方面的思考，基于题目反应理论（IRT）的题目功能差异（DIF）研究，基于 WEB 的 HSK 动态作文语料库的建设，网上模拟 HSK 考试系统和练习系统的开发，题库中等值试卷生成方式的改进，关于国际汉语教师能力考试的探索和思考，将规则空间模型应用于语言测试的尝试，等等。从本辑中也可以反映出我们在 HSK、MHK、ZHC、公务员考试、国际汉语教师能力考试等各个领域中所进行的努力。

“以文会友”，我们希望通过文集与同行们展开学术交流，互相启发。2003 年谢小庆在《湖北招生考试》发表了《考试应体现谁的意志》一文。“一石激起千层浪”，引发了许多同行撰文进行商榷。2006 年，湖北人民出版社出版了由罗金远同志主编的《考试应体现谁的意志》一书，收入了 20 多篇参加讨论的论文。从这场讨论中，许多人都加深了对考试问题的理解。在本辑中，我们也谈出了一些自己独到的看法。我们期待着这些看法也能激发起同行们的讨论愿望。

编者

2008 年 2 月 26 日

目　录

概论

实证研究

汉语水平考试（HSK）

中国少数民族汉语水平等级考试（MHK）

国家职业汉语能力测试（ZHC）

汉语作为外语教学能力认定考试（TCSOL）

文献综述

国家职业汉语能力测试（ZHC）

汉语作为外语教学能力认定考试（TCSOL）

概　论

美国1999年版与1985年版《教育与心理测试标准》的对比分析*

谢小庆

摘要：《教育与心理测试标准》是教育与心理测验领域中的一份权威性文献，它同样也是考试领域中的行业标准，是考试工作者和考试研究人员的行为准则。该标准在1999年进行了修订，已由美国教育研究协会、美国心理学会和美国国家教育测量学会共同颁布。本文将1985年与1999年标准进行了对比与分析，讨论了1999年版标准的特点及其颁布对我国测验研究工作的意义。

关键词：《教育与心理测试标准》　对比　分析

---

* 原载《中国考试》2004年第4期。

1999年美国教育研究协会（American Educational Research Association，AERA）、美国心理学会（American Psychological Association，APA）和美国国家教育测量学会（National Council on Measurement in Education，NCME）三家共同颁布了新版的《教育与心理测验标准》（Standards for Educational and Psychological Testing，以下简称《标准》）。《标准》一书共15章，包含了效度，信度和测量误差，测验的开发和修订，评分，常模建立，分数等值，分数报告，公平性，考试有关各方（考试开发者、考试使用者、考生）的责任与义务，考试在教育、医疗和职业领域中的应用等内容。《标准》系统阐述了考试编制和实施所应遵循的基本原则，规定了考试所应达到的技术要求。《标准》是教育与心理测验领域中的一份权威性文献，体现了考试领域中的行业标准，是考试工作者和考试研究人员的一本重要的参考书。

一、《标准》的历史沿革

在1999年之前，颁布《标准》的3家机构曾经颁布过5个有关测验开发和使用的文件。第一个是1954年由APA颁布的《关于心理测验和诊断技术的技术建议》。第二个是1955年由国家教育协会颁布、由AERA和NCME编制的《关于成就测验的技术建议》。第三个是1966年由APA出版、由APA、AERA和NCME共同编制的《教育与心理测验及手册的标准》。第四个是3家机构于1974年对第三个文件的修订版。

1977年，颁布《标准》的3家机构成立了一个联合委员会来对1974年版《标准》进行审查。委员会包括12名委员，主席是诺维克（M. R. Novick），副主席是R. L. Linn和S. W. Sherman。联合委员会提出了对《标准》进行修订的一些指导性意见和基本的工作原则，主要包括：

1. 内容要覆盖到测验在各个方面的应用；

2. 是一个关于规范专业实践技术标准的说明，而不是一个社会行为法规；

3. 在判断测验技术的充分性、测验使用的适当性、测验分数的合理性方面，这一文件应提供依据；

4. 要求测验的编制者、出版者和使用者收集并提供充分的资料，使具备资格的审查者可以据此判断该测验是否达到了使用标准；

5. 尽管《标准》本身并不具备强制机制，但《标准》应体现强烈的道义力量；

6. 认识到并非所有的标准都适用于内容广泛的各种测验和测验的使用情境；

7. 其表达方式应尽量使多数使用测验和测验分数的人都可以理解；

8. 并不禁止在测验的编制、使用和解释方面进行新的探索；

9. 反映权威专家们今天的共识；

联合委员会经过8年的工作，于1985年颁布了《标准》的第5个版本。

1991年，颁布《标准》的3家机构开始酝酿对1985年版本进行修订。1993年11月正式成立了修订《标准》的联合委员会并召开了第一次会议。委员会有15名成员，由E. Baker和P. Sackett共同担任主席。

《标准》修订的未定稿曾三次大面积分发，广泛征求意见。委员会共收到来自74家机构的近8000页的评论意见。这些机构不仅包括教育部提高教育质量办公室全国教育统计中心、人事管理总署人力资源及开发中心、劳工部就业培训管理局、司法部移民归化局、国防部助理部长办公室、美国平等就业机会委员会等政府机构，而且包括美国高等教育协会、美国医疗专科理事会等行业协会，包括全美医师执照测验委员会、全美注册心理咨询师委员会等资格认证机构，另外还包括大学委员会、教育测试服务中心（ETS）、

美国大学考试中心（ACT）等民间考试机构。

二、3家颁布机构对《标准》1999年版的批准意见

经过广泛吸收和听取有关机构的修改意见，联合委员会最终完成了《标准》的修订工作。3家颁布机构分别根据自己的审批机制和程序批准了经过修订的《标准》。

美国教育研究协会的批准意见是："批准本《标准》的修订完成表明，原则上我们相信《标准》代表了目前本领域公认的专家们就测量实践中应遵循的准则所达成的共识。测验的研制、施行、出版发行及使用各方人士，均须照此办理。"

美国心理学协会的批准意见是："《标准》的通过表明本协会将采用该文献作为准则。"

全美教育测量学会的批准意见是："本学会批准《标准》的修订完成，相信这些标准将有助于更合理和更负责任的测试实践。此项对《标准》的批准意味着本学会会员在工作中应尊重这些《标准》，这种尊重是一种专业责任。"

三、新《标准》的特点

与1985年版《标准》相比，1999年版《标准》具有以下一些新的特点。

（一）内容的增加

在新版本中，增加了许多新的内容，反映了20世纪90年代测验领域的新发展。1985年版本100页，1999年版本为194页，篇

幅大幅度增加。1985年版本共包含“标准”180条，新版本包含“标准”264条。在《标准》中包含一个名词术语解释。在1985年版本中，包含术语解释122条。在新版本中，包含术语解释199条，增加了77条，其中包括信度效度系数的调整、分析性记分、偏见、（参数）标定、项目功能差异、经典测验理论、项目反应理论、高利害测验、测验信息函数等许多重要的术语。在新版本中，增加了许多新的内容，例如，对效度的重新定义和分类，对公平性问题的深入探讨，对题目反应理论（IRT）的介绍等。

（二）对效度的重新定义和分类

与1985年版本相比，1999年颁布的《标准》最突出的特点是重新定义了测验效度。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）3种，效度也被相应地划分为3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。

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提出构念效度（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明显占据了上风，构念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年版《标准》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）

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，包括基于内容的证据（evidence based on content）、基于反应过程的证据（evidence based on response processes）、基于内部结构的证据（evidence based on internal structure）、基于与其他变量之间关系的证据（evidence based on relations to other variables）和基于测验结果的证据（evidence based on consequences of testing）。

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。

（三）对公平性问题更多的关注

在1985年版本中包括4个部分共16章内容。4个部分是测验编制和评价的技术标准（含效度、信度等5章）、测验使用的专业标准（含临床测验、学校测验、雇用测验、证书测验等7章）、特殊应用的标准（含对少数民族和对残疾人的测验2章）和施测过程的标准（含施测、计分、分数报告等2章）。

在1999年版本中包括3个部分共15章内容。3个部分是编制、评价和文件准备（含效度、信度等6章）、测验公平（含测验公平、考生权利、少数民族、残疾人等4章）和测验应用（含心理测验、教育测验、职业测验等5章）。在新版本中，“测验公平”单独成为3个部分中的一个部分。

在1985年版本中，虽然在一些地方提及公平性问题，但并没有专门章节的深入探讨。在1985年版本中，认为如果一项考试不会系统地高估或低估某一特定考生群组，这个考试就是公平的（第12页）。新版本超越了关于公平性的这种定义，对“公平”概念的不同含义进行了更深入的探讨，分别讨论了“没有偏见的公平”、“考生受到同等对待的公平”、“学习机会的公平”、“考试结果的公平”等不同的公平概念，分别探讨了可能导致不公平的不同因素，如来自测验内容的因素和来自解题过程的因素等。