考试研究文集

谢小庆　杨　洋／主编

经济科学出版社

考试研究文集

（第6辑）

谢小庆　杨　洋　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6辑）/谢小庆，杨洋主编. —北京：经济科学出版社，2011.4.

ISBN 978-7-5141-0468-4

Ⅰ.①考… Ⅱ.①谢…②杨… Ⅲ.①考试学-文集 Ⅳ.①G424.74-53

中国版本图书馆CIP数据核字（2011）第035888号

责任编辑：卢元孝
责任校对：徐领柱
版式设计：齐 杰
技术编辑：王世伟

考试研究文集（第6辑）

谢小庆 杨 洋 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：100142

总编部电话：88191217 发行部电话：88191540

网址：www.esp.com.cn

电子邮件：esp@esp.com.cn

北京三木印刷有限公司印装

880×1230 32开 10印张 260000字

2011年4月第1版 2011年4月第1次印刷

ISBN 978-7-5141-0468-4 定价：25.00元

前　言

据上海《新闻晨报》2010 年 11 月 16 日（A22 版）报道，iphone4 手机在市场上非常紧俏。从苹果官方网站上可以订购，但经常无货。于是“黄牛党”们日夜蹲守在苹果网站上抢购。一旦得手，一台转手就可获利千元。

2010 年，苹果公司的 ipad，更是一机难求。靠 ipod、iphone 和 ipad 赚得盆满钵满的苹果公司，让不知多少中国公司嫉妒得眼中冒火。不知多少中国公司，为了销售自己的产品而不惜血本砸广告，为了销售自己的产品而使尽浑身解数。

我看到，在 iphone 和 ipad 上写着：“设计于加利福尼亚，组装于中国”。在世界上被抢购的 iphone 和 ipad，大多是在中国组装的，大多出于“富士康”的工厂。

如果需要为 2010 年确定几个关键词，那么，“富士康”可能要在候选者之列。2010 年，“富士康”的“N 连跳”曾经成为媒体关注的焦点。“富士康”是最成功的华人企业之一，“富士康”的老板郭台铭也是世界上最富有的华人之一。

“富士康”属于高科技企业，但它却没有自己的核心技术，它仅仅是苹果等真正高科技企业的组装厂。在苹果产品的利润中，“富士康”的所得不过百分之几，非常微薄。

几乎每一个 iphone 和 ipad 的玩家，都会被其产品所包含的想象力所折服。苹果公司靠什么赚得盆满钵满？显然，靠的是想象力。中国人在付出了沉重的资源、环境以至“N 连跳”的代价之后，为什么只能分到一点薄利？显然，是由于想象力和创造力不足。

中国有 13 亿人，中国人的想象力和创造力哪里去了呢？只要

稍微去关注一下我们的小学和中学，就不难发现，中国人的创造力和想象力在童年时代就被“应试教育”摧残了，吞食了。

新中国成立已经61年。61年来，不知多少中华民族最优秀的人才从北京大学、清华大学、复旦大学的物理系、化学系和生物系毕业。至今，这些毕业生中尚没有产生出一个诺贝尔奖的获得者。

有人辩解说，不是我们的学校培养不出人才，而是我们尚缺乏优秀人才做出发明创造的科研条件。这种说法是站不住的。从20世纪80年代初开始北大、清华和复旦就开始向欧美输送毕业生，近30年来，不知多少北大、清华和复旦的毕业生进入欧美的研究机构，享受了很好的研究条件。30年过去了，并没有产生一个诺贝尔奖得主。显然，不能把不出人才归因于缺乏研究条件。

我认为，我们的学校培养不出一流人才，主要的原因是应试教育，是由于孩子们的好奇心和创造力被应试教育所破坏，是由于学生们受到了“童子伤”。

可喜的是，今天已经有越来越多的人认识到这一点。2009年9月4日温家宝同志在北京三十五中听课时说：“建国以来培养的人才尤其是杰出人才，确实不能满足国家的需要，还不能说在世界上占到应有的地位……我们出去这么多留学生，也成长了一批人才，充实了各行各业，但确实很少有像李四光、钱学森、钱三强那样的世界著名人才。每每想到这些，我又感到很内疚……我们在过去相当长的一段时间里比较重视认知教育和应试的教学方法，而相对忽视对学生独立思考和创造能力的培养。”（见2009年10月12日《人民日报》）

我们历来认为，之所以存在“应试教育”与“素质教育”的对立，就是因为考试所考查的不是“素质”而是“背书”。一旦考试可以考查“素质”，应试教育与素质教育就得到了统一。中国是一个考试大国，各种各样的考试五花八门，但真正能够促进素质教育的考试却难得一见。作为职业的教育与心理测量研究人员，作为职业的考试工作者，我们有责任推动考试的科学化，有责任开发更

多可以考查“素质”而不仅仅是考查“背书”的好的考试，有责任促进素质教育与应试教育的统一。

在这一辑中，记录了我们一年多以来的探索足迹。内容包括关于教育思想和教材教法的讨论，关于标准化考试的思考，关于能力测验开发路线图的思考，关于高考改革的思考，关于考试公平的讨论，关于现代测验理论项目反应理论（IRT）的讨论。内容涉及融合模型等认知诊断模型、聚类方法、多元非线性回归模型，等等。内容还涉及作文自动评分、集团作弊甄别、汉字应用水平测试、国际汉语教师能力认定考试、公务员录用考试等考试应用领域。

2010年2月3日，胡锦涛同志在省部级主要领导干部加快经济发展方式转变专题研讨班上说：“国际金融危机使我国转变经济发展方式问题更加凸显出来，国际金融危机对我国经济的冲击表面上是对经济增长速度的冲击，实质上是对经济发展方式的冲击。综合判断国际国内经济形势，转变经济发展方式已刻不容缓。”

2010年2月4日，温家宝同志在相同的研讨班上说：“发展科技、教育和文化事业，全面提高人的素质，是转变经济发展方式、实现可持续发展的关键。”

我们认为，“转变经济发展方式”确实已经“刻不容缓”，而其“关键”，确实是“提高人的素质”。这里，奉上我们的一些研究和思考，就教于同行们。我们希望通过交流与讨论，与同行们共同推动考试改革，推动教育改革，为我国转变经济发展方式，准备具有创造力的人才。

编者

2010年11月23日

目　录

概　论

实证研究

文献综述

调查报告

其　他

概　论

谈语言能力测验开发的路线图

谢小庆
（北京语言大学）

【摘要】本文结合“汉字应用水平测试（HZC）”的开发实践讨论了语言能力测验开发的路线图。本文认为，在语言能力测验的开发中，在许多涉及复杂心理特征的能力测验和职业测验的开发中，理论上似乎合理的“观察—归纳”路线图，在实践中是行不通的。在这些测验的开发中，我们往往需要采用“猜测—反驳”的路线图。

【关键词】测验；考试；语言测试

按照通常的观点，在开发一个语言能力测验之前，首先需要对语言活动进行调查，界定测验所要测量的“构念（construct）”，回答“考什么”的问题。之后，对具有不同语言水平的考生的行为表现进行描述，制定出基于“能做（can do）”的语言能力等级标准，并给出必要的任务举例。例如，教育部和国家语言文字工作委员会2006年联合颁布的《汉字应用水平等级及测试大纲》（以下简称《测试大纲》）中，对各个等级的标准进行了能力描述。在国家汉语国际推广领导小组办公室2007年公布的《国际汉语能力标准》中，不仅对各个等级的标准进行了能力描述，还给出了各个等级标

准的任务举例。理论上，《测试大纲》是语言能力测验开发的依据，测验编制应以这些关于各个级别的能力描述为依据。

经过多年的语言测试实践，我发现，这种“行为分析—构念界定—能力描述—测验编制”的路线图是行不通的。

我们以教育部和国家语言文字工作委员会主持开发的“汉字应用水平测试（HZC）”为例来进行讨论。

在《测试大纲》中，对汉字应用水平从三个方面进行了能力描述：第一，所掌握字量；第二，对字形、字音、字义的辨识能力和使用水平；第三，阅读和书写水平。关于字量，大纲的编写者认为：“3500 常用字是……中等偏下文化水平人群的识字量……根据我们的实际测查，具有高中文化程度的人群一般识字量为 4000 字左右……具有大学文化程度的人群一般识字量为 4500 字左右……具有大学以上文化水平、从事文字工作、且具有较高汉字应用水平的人群的识字量为 5500 字左右。”（孙曼均，2004）

这一研究结果可以受到质疑。根据大纲编写者提供的资料，李白 994 首 7.7 万字的诗文用字 3560 个，杜甫 1500 余首诗用字 4350 个，白居易 3000 余首诗共 18 万字，用字 4600 个。66 万字的《毛泽东选集》1～4 卷的用字量为 2891 个，毛泽东公开出版的全部著作也仅用单字 3136 个。（孙曼均，2004）按照“高中生识字 4000 字”的观点，高中生读毛泽东的著作应几乎不会遇到不认识的字，读李白的诗应基本不会遇到不认识的字。按照“大学生识字 4500 个”的观点，大学生读杜甫和白居易的诗应基本不会遇到不认识的字。这可能与实际情况有很大距离。我与多位毕生从事语言文字研究工作的我国顶级语言学家们讨论过识字量的问题，多数专家认为普通人的识字量在 3000 左右。

在由教育部颁布的基础教育新课程改革的指导性文件《全日制义务教育语文课程标准》中，要求初中毕业生认识常用汉字 3500 个，其中 3000 个左右会写。在《普通高中语文课程标准》中，并没有再提出字量方面更高的要求。

我认为，高中语文课程标准的制定者们没有在识字量方面提出更高的要求是非常正确的。这样规定，遵循了语言能力发展的规律，符合汉语的特点，符合实际情况。事实上，对于已经完成高中教育的人群来说。他们“汉字应用水平”的差异主要不是体现在掌握字量的多少上，而是体现在对有限数量的汉字的应用上，体现在对汉字义项掌握的多寡上，体现在根据一定的语境选用适当的汉字上，体现在汉语表达的准确、得体和优雅上，体现在与汉字应用有关的语感上。

在《测试大纲》中包含有《汉字应用水平测试字表》，总字量为5500字，分甲、乙、丙三个子表，甲表为4000字，乙表为500字，丙表为1000字。字表按字形来计算汉字数量。实际上，许多具有相同字形的汉字具有多个不同的读音，具有多个不同的义项，同一个字，不同的读音，不同的义项之间难度差异很大。具有不同汉字应用水平的人在汉字读音和汉字义项的掌握上，差异很大。例如，“数”字有 shǔ、shù、shuò 三个不同的读音，不同的读音之间具有明显的难度差别。又如“马前卒”中的“卒”，“生卒”中的“卒”，“卒业”中的“卒”，“卒中（读 cùzhòng，一种疾病名称）”中的“卒”，具有不同的含义，不同的含义之间具有明显的难度差别。

从理论上说，可以通过调查来了解不同人群的识字量，在此基础上确定 HZC 不同级别的识字量标准。根据标准，编制汉字应用水平测验。《测试大纲》编写者的思路是：对高中毕业生、大学毕业生、从事语言文字工作的专业人员这三个群体进行调查，了解他们的平均识字量，以“观察—归纳”方式得到各个群体的平均识字量，以此作为三级、二级和一级的标准。实际上，这种“观察—归纳”的路线图是行不通的。不用说了解特定人群的平均识字量，就是仅仅了解一个人的识字量，了解一个人从字音、字形、字义、书写等几个方面可以掌握的汉字数量，也是很困难的。关于“识字”的标准就很难把握。何谓“识字”？可以是能认能写，也可以是能

认不能写；可以是掌握多音字的一种发音，也可以是掌握多音字的多种发音；可以是掌握多义字的一种义项，也可以是掌握多义字的多种义项；可以是仅仅正确地认读和书写“辩”和“辨”字，也可以是准确地把握“分辨（分辨方向）”和“分辩（不容分辩）”之间的微妙使用差异。

为了了解一个人或一群人的识字量，首先需要有一个可靠、有效的测量工具。如果我们已经具备了这样的一个测量工具，就没有必要再去开发 HZC，再去为 HZC 确定各个标准的达标分数线。这里，陷入了逻辑循环。

正是由于所采用的测量方法缺乏可靠性，正是由于所采用的识字量测量工具缺乏效度和信度，《测试大纲》编写者关于“具有大学文化程度的人群一般识字量为4500字左右”（孙曼均，2004）的结论是不可信的。

就科学研究的一般方法而言，通常有两种思路：一种是“观察—归纳”的思路，即不带偏见地对所有可能的相关因素进行客观观察，在全面、客观地掌握有关事实的基础之上归纳出规律性的结论；另一种是“猜测—反驳”的思路，根据研究者对所研究问题的已有理解和经验直觉，大胆地提出假设，设计试验条件对这种假设进行检验（反驳），在检验过程中对假设进行修正或放弃假设。研究相对比较简单、影响因素有限的问题时，“观察—归纳”的方法可能是有效的。但是，在研究相对比较复杂、影响因素很多的问题时，则往往需要采用“猜测—反驳”的方法。

即使是在物理学研究中，“观察—归纳”的方法的应用范围也是很有限的，也只能应用于有限的中观物理现象研究，在关于微观和宏观的物理学研究中，也常常是行不通的（谢小庆，1985，此文第一节的标题是“对传统‘观察—归纳’方法的物理学批判”）。早在19世纪，天才的思想家恩格斯就已经认识到这种“观察—归纳”方法的局限性。他说：“按照归纳派的意见，归纳法是不会错误的方法，但事实上它是很不中用的，甚至它的似乎最可靠的结

果，每天都被新的发现所推翻。”（恩格斯）

为“汉字应用水平”这样较复杂的心理属性建立标准，曾经在物理学研究中取得过一些成绩的“观察—归纳”方法在这里是行不通的。此时，“观察—归纳”方法面临逻辑悖论：测量标准（量表）的建立首先需要有测量工具，而测量工具的研制又需要以建立测量标准（量表）为前提。

实际上，为了建立“汉字应用水平”的等级标准，我们需要另一种路线图，即“猜测—反驳（检验）”的路线图。我们需要首先从我们关于汉字应用的经验出发，根据我们关于汉字应用水平的直觉，编制出一个汉字应用水平测验，开发出一个测量工具。我们大胆猜测这是一个有效的工具。之后，我们小心翼翼地对这一工具进行检验，广泛收集效度资料，比较不同学历、不同职业、不同专业、不同性别、不同地区、不同年龄、不同类别的学校（211、一本、二本、高专、高职等）的应考者的成绩，根据收集到的效度资料不断对测验进行修订，删除那些效度不好的题目。在获得了一个比较好的测量工具以后，根据这一测验的分数来建立标准，例如，达到200分为三级，达到400分为二级，达到600分为一级。这就是2007年、2008年HZC试测和分数报告的实际过程，这就是根据2008年的HZC试测结果建立HZC分数体系和等级标准的实际过程。

在建立了量表之后，我们可以根据效度资料对分数做出更丰富的解释，例如，高中毕业生的平均水平为500分，大学毕业生的平均水平是600分，中学教师的平均水平是650分，而中学语文教师的平均水平是700分，等等。

与HZC相似，在许多能力测验和职业测验的开发中，工作分析的思路也是行不通的。例如，当笔者1988年着手开发用于公务员录用考试的“行政职业能力测验”的时候，我们曾经有一个清楚的“观察—归纳”的路线图：对政府工作人员进行工作分析—归纳政府工作人员需要的知识和能力—根据工作分析结果编制测验。事实上，我们曾经在当时的国家建材局进行了比较规范的工作分析研

究。其后，我们也多次进行关于政府工作人员的工作分析。经过多次尝试，我们逐渐认识到，由于政府工作的复杂性，虽然我们可以通过工作分析得到一些关于政府工作人员所从事活动的描述，可以得到一些关于从事政府工作所需要的知识、能力、心理特点的资料，但是，这些工作分析结果很难与测验编制挂钩，在工作分析结果和测验题目之间存在着难以跨越的鸿沟。在用于公务员录用考试的“行政职业能力测验”的开发中，我们实际采用的是一条“猜测—反驳”的路线图：根据关于政府工作人员所需要能力的经验直觉尝试性编制测验—在测验的施测过程中不断收集效度资料—根据收集到的效度资料来对测验进行改进和调整。

又如，当笔者受原人事部人事考试中心（现为人力资源和社会保障部人事考试中心）委托开发用于企业管理人员选拔的“企业管理职业能力倾向测验”的时候，也曾试图按照“工作分析—测验编制”的路线图进行测验开发工作，曾经采用了世界上权威的《工作分析问卷（PAQ）》和《专业和管理职位分析问卷（PMPQ）》进行工作分析。在进行了一系列的工作分析尝试之后，最终发现，由于企业管理人员工作性质的复杂，很难将工作分析的结果与测验的编制相联系。另一方面，关于测验的效度研究却为测验的改进和完善提供了许多实证依据。针对不同测验内容，针对所考察的不同能力，针对不同的题型，我们通过多种渠道收集效度资料。我们将测验分别施测于企业中的管理人员和非管理人员，分别施测于企业中较成功的管理人员和不太成功的管理人员，分别施测于企业中的管理人员和其他组织（机关、学校）的管理人员，对施测结果进行分析，考察不同分测验、不同测验内容、不同题型的效度。（谢小庆，1999）

在受原人事部人事考试中心委托进行的关于经济师、经济员任职资格考试的研究中，我们也感受到基于工作分析的“观察—归纳”路线图的局限性。（王二平、谢小庆，1994）

综上所述，在语言能力测验的开发中，在许多涉及复杂心理特

征的能力测验和职业测验的开发中，理论上似乎合理的“观察—归纳”路线图在实践中是行不通的。在这些测验的开发中，我们往往需要采用“猜测—反驳”的路线图。

参考文献

恩格斯：《自然辩证法》，人民出版社 1971 年版，第 206 页。

国家汉语国际推广领导小组办公室：《国际汉语能力标准》，外语教学与研究出版社 2007 年版。

教育部、国家语言文字工作委员会：《汉字应用水平等级及测试大纲》，广东教育出版社 2006 年版。

孙曼均等：《汉字应用水平测试用字的统计与分级》，载《语言文字应用》2004 年第 1 期，第 68～69 页。

王二平、谢小庆：《银行保险业经济系列专业职称资格的职务分析》，载《心理学动态》1994 年第 1 期。

谢小庆：《现代心理学研究成果的认识论意义》，载《中国社会科学》1985 年第 1 期，第 110～124 页。

谢小庆等：《用于企业人事管理的〈企业管理能力倾向测验〉》，载《心理学报》1999 年第 31 卷第 2 期，第 222～229 页。

中华人民共和国教育部：《全日制义务教育语文课程标准》，北京师范大学出版社 2001 年版。

中华人民共和国教育部：《普通高中语文课程标准》，人民教育出版社 2003 年版。

The Approaches to a Language Competence Test

Xie Xiaoqing

(Beijing Language and Culture University)

Abstract: According the practice of development of Chinese Character

Application Proficiency Test (HZC) this paper discussed the approaches to develop a language proficiency test. Argued that we have to develop competence test along the "conjecture-refutation" approaches rather than the "observation-induction" approaches, though later looks theoretically reasonable.

Key words: test; examination; language testing

（原刊于《考试研究》2010年第1期）