标准化英语考试等值可行性研究

BIAOZHUNHUA YINGYU KAOSHI DENGZHI KEXINGXING YANJIU

吕剑涛　著

人民出版社

标准化英语考试等值可行性研究

BIAOZHUNHUA YINGYU KAOSHI DENGZHI KEXINGXING YANJIU

吕剑涛　著

人民出版社

责任编辑:李椒元
装帧设计:徐　晖
责任校对:吕　飞

图书在版编目(CIP)数据

标准化英语考试等值可行性研究/吕剑涛 著. —北京:人民出版社,2017.9
ISBN 978-7-01-017541-6

Ⅰ.①标…　Ⅱ.①吕…　Ⅲ.①英语-标准化考试-研究-中国
Ⅳ.①H310.42

中国版本图书馆 CIP 数据核字(2017)第 063003 号

标准化英语考试等值可行性研究
BIAOZHUNHUA YINGYU KAOSHI DENGZHI KEXINGXING YANJIU

吕剑涛　著

人民出版社 出版发行
(100706　北京市东城区隆福寺街 99 号)

北京毅峰迅捷印刷有限公司印刷　新华书店经销

2017 年 9 月第 1 版　2017 年 9 月北京第 1 次印刷
开本:710 毫米×1000 毫米 1/16　印张:18
字数:269 千字　印数:0,001-3,000 册

ISBN 978-7-01-017541-6　定价:40.00 元

邮购地址 100706　北京市东城区隆福寺街 99 号
人民东方图书销售中心　电话 (010)65250042　65289539

本著作由广东外语外贸大学人文社科重点研究基地“大学英语能力测试与评估研究中心”资助出版

该著作为本人攻读博士研究生期间成果。衷心感谢我的导师刘建达老师在我攻读博士学位期间对我科研工作的悉心指导。刘老师使我从零慢慢掌握语言测试领域理论及研究方法，给予我许多从事科学研究的机会。在此再次感谢刘老师对我的教导、关爱和提携。

感谢桂诗春教授对本研究技术攻关上的帮助。桂老师在我的论文研究过程中向我提供了大量相关的参考文献，并多次邀请我到其家中与我畅谈自己多年从事语言测试科研工作的经历，使我从中受到启发，获益良多。桂老师在本论文的试题质量统计分析、模型参数校准等方面提供了技术指导。

本研究还得到了亓鲁霞老师和金艳老师的鼓励和支持，得到广西壮族自治区招生考试院命题处的大力协助，在此一并感谢！

我要感谢我的家人在我读博期间对我的谅解和关心。我要感谢、怀念一直支持我攻读博士研究生，于 2012 年离世的父亲。

目　录

引　言

当前在教育部统一领导和部署下，全国外语改革尤其是英语改革已经全面展开，其中关于英语考试的改革显得尤为突出。2014 年 9 月 3 日国务院颁布的《关于深化考试招生制度改革的实施意见》明确指出，加强外语能力测评体系建设。该测评体系是覆盖大中小学各教育阶段、覆盖听说读写译综合能力、覆盖外语学习及教学与测评的评价系统。其中英语能力等级量表的研制已经初步完成，逐步开始进入实证研究阶段。而大规模的标准化考试是保障此项工作顺利完成的重要环节，也是未来在新的评价体系下开展英语等级考试的重要手段，以促进我国英语考试与国际成熟考试（如雅思、托福等）的顺利接轨，为我国的外语学习者、学校及社会提供服务。

设立能力考试（proficiency test）的主要目的之一是进行人才选拔（Bachman，1999）。为了兼顾公平性，应试者需要在尽可能相同的条件下参加考试（桂诗春，2000a），所以能力考试多为标准化考试。目前，我国设立了多个英语标准化考试。这些考试一般以客观题为主，从听力、词汇、语法、阅读多个方面考查应试者的英语能力水平。每年，参加这些考试的人数众多，举办考试需要耗费大量的人力物力。目前国内较大规模的标准化英语考试为全国普通高等学校招生统一考试（英语高考）（National Matriculation English Test，NMET）。该考试成绩作为各高校录取应届高中毕业生的依据，全国每年大概有 1000 万人左右参加①。考试期间甚至在考点附近路段实行封闭式交通管

① 统计数据来源于财新网（http://special.caixin.com/2013-05-14/100527558.html，最后访问网页时间：2013/9/6）。

制或车辆行驶限制。该考试第一部分为听力客观题;第二部分为英语知识运用,全部为客观题,考查考生词汇、语法知识(单项填空)和语言综合应用能力(完形填空);第三部分为阅读理解,大部分为客观题。另一较大规模的标准化考试是全国大学英语四、六级考试(College English Test Band 4 and Band 6, CET)。近年来每年参考人数超过1100万人①,是全球参考人数最多的考试,已成为各级人事部门录用大学毕业生的标准之一②。考试第三部分为听力理解,大部分为客观题;第四部分为深度阅读理解,部分为客观题;第五部分为完形填空,全部为客观题。由于参与考试的人数众多,而且在很大程度上决定着考生今后接受教育和参加工作的机会,所以属于高风险考试(high-stakes exams)。这些考试对英语学习、教学,甚至人们日常生活造成一定影响,并产生了一些矛盾和冲突。

目前国内举行的英语标准化考试有一个共同特点:考试是一次性的。举办方每年在同一时间使用一套试卷举行考试。虽然该方法很好地满足了考试的公平性,但同时也给应试者带来一定的压力。应试者担心在单次考试中发挥不佳,在考试前容易产生紧张、焦虑、恐惧等负面心理。其实,无论是何种类型的考试,都应该使应试者怀着轻松、愉快的心情参加(Bachman,1999)。因此,标准化考试除了需要兼顾公平性外,还应该给予考生更多参与考试的机会。许多教育工作者主张在一年内的多个时间段设立多场次考试,以减轻单次考试对应试者带来的压力,即所谓"一年多考"。

我国在2010年将"一年多考"提上了教育改革的议程③④。实施"一年多考"的优势是,应试者可以根据个人情况选择适合自己的考试场次。实施多场次考试为考生提供更多参加考试的机会,但同时也产生另一实际问题,应试者的成绩隐含一定的随机性。

目前,国内标准化考试的成绩一般使用原始分公布。考生的原始得分很

---

① 近年来参加全国大学英语四、六级考试的人数为1800万左右,单次考试在900万左右。

② 该消息来源于外语教育网(http://www.for68.com/zhinan/CET/,最后点击时间:2013/9/7)。

③ 见《国家中长期教育改革和发展规划纲要(2010—2020)》。

④ "一年多考"改革不只受到教育部关注,也是国务院指出要在近期实现的举措,特别是三中全会以后,"一年多考"已作为一项要加快实施执行的政策。

大程度上受试卷难度影响：试卷难度越低，分数越高；难度越高，分数越低。换言之，同一考试不同试卷的分数是不能直接比较的。为了使不同试卷的分数具有可比性，国内某些标准化考试，如全国大学英语四、六级考试对不同年份施测的试卷进行了等值处理。同理，为避免试卷间难度差异对考试成绩造成的随机性，实行“一年多考”亦要对不同场次施测的试卷实行等值。

试卷等值的根本目的是根据试卷间的难度差异调整考生成绩（Kolen & Brennan，2004）。国内在 90 年代初已有讨论如何等值标准化考试试卷的文献（桂诗春，1989；1990a；1990b；1990c）。近年来也有文献探讨如何用不同方法进行试卷等值（朱正才，2005；朱正才等，2003）。但这些研究都是以等值不同年份试卷为研究背景。国外已有大量文献讨论使用不同方法进行试卷等值的可行性（Baker & Al-Karni，1991；Beard & Pettie，1979；He et al.，2012；Kolen & Whitney，1981；Lee et al.，1998；Livingston et al.，1990；Morrison & Fitzpatrick，1992；Petersen et al.，1983；Schmitt et al.，1990；Sontag，1984；Wright，1995；Zhang，2010），也有文献详细介绍试卷等值的理论和操作（Kolen & Brennan，2004；von Davier，2011），但从实施“一年多考”角度探讨试卷等值的文献还不多见。实施“一年多考”的关键问题在于如何校准和等值出数量较多的试题。经过校准和等值的试题可以为多场次考试组合出多套平行试卷。

试题难度是试题校准的主要对象。经典测试理论中，试题难度由试题通过率反映。试题通过率除了由试题自身难度决定外，还受考生能力水平影响（Fox & Jones，1998）。在项目反应理论（Item Response Theory，IRT）下，应试者答对试题的概率取决于两个因素：应试者的潜在特质（latent traits）和试题的特点（Hambleton et al.，1991），试题特征参数值独立于考生样本（Hambleton & Swaminathan，1985；Lord，1980），该特点使 IRT 理论广泛应用于题库建设中（Jones，1992；Szabó，2008；桂诗春，1989；1990；余民宁，1993c）。

IRT 理论中使用较多的模型有：Rasch 模型（Wright & Stone，1979）、二参模型（two-parameter model）（Birnbaum，1968）和三参模型（three-parameter model）（Lord，1980）。三个模型对试题特征考虑的因素有所不同：Rasch 模型只考虑试题难度；二参模型考虑试题难度和区分度；三参模型除考虑试题难度和区分度外，还考虑试题猜测度。丹麦数学家 Georg Rasch 发现 Rasch 模型下

的参数估计值能真正独立于样本。Georg Rasch 首先将项目反应理论应用到心理测验开发中,他发现"(Rasch 模型使)不依赖样本估计出试题难度,不依靠某道试题估计出个人能力的问题得到解决"(1960)。Rasch 模型可以看作项目反应理论的二参、三参模型的简化模型,即假设试题猜测度为零,试题区分度为 1。在 Rasch 模型下,正确回答试题的概率由试题难度和考生能力共同决定,即:

$$p_j(\theta_i) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)} \tag{1.1}$$

其中,θ_i为考生 i 的能力水平估计值,b_j为试题 j 的难度。

使用 Rasch 模型有两个优势:一是可以用小至 100 的考生样本校准试题(Wright,1977)。使用其它更复杂的模型时,需要 500(二参模型)甚至 1000(三参模型)的考生样本(Hambleton,1989);二是使用 Rasch 模型能估计出充分一致的试题难度值(Al-Owidha,2007;Slinde & Linn,1978)。Szabó(2008)指出使用 Rasch 模型建设题库的三个理由:第一,用于校准试题难度值的样本数往往比较小,不适宜使用更复杂的二参、三参模型;第二,二参、三参模型的参数估计运算较复杂,参数估计值往往难以获取;第三,在所有 IRT 模型中,只有 Rasch 模型估计的参数值在理论上真正独立于考生样本,只有 Rasch 模型能反映出试题间的绝对难度差异。Rasch 模型在题库建设中的可行性已被许多研究证明(Henning,1986;Jones,1992;Szabó,2008)。

目前学界对 Rasch 模型的适用性仍存在争议。研究表明,Rasch 模型与数据的拟合程度往往不如二参、三参模型(Al-Owidha,2007;Choi & Bachman,1992)。有的学者指出 Rasch 模型的等区分度和最小猜测度假设不符合试题的实际情况:试题区分度不可能完全一致(Slinde & Linn,1978),而且试题通常存在一定的猜测因素(Skaggs & Lissitz,1986)。以上对 Rasch 模型的种种质疑使模型选择变得棘手:一方面,Rasch 模型简化了应试者答题行为的产生机制,比较有利于对试题和考生行为作研究;另一方面,Rasch 模型的高度理想化特性又有悖于试题实际情况之嫌,由更符合试题实际情况的二参、三参模型校准出的试题参数应该比 Rasch 模型校准出的试题参数更可靠、准确。那么,校准英语客观试题究竟应该选择 Rasch 模型,还是二参、三参模型?目前学界

对此并无定论。

值得指出的是，Rasch 模型与早期心理测试理论的初衷是一致的。早期的心理测试理论只希望建立一个关于试题难度较稳定的量度，使其不随应试者能力的变化而变化（Gulliksen，1950）。如果再额外考虑试题的其它特征（区分度、猜测度等），会使试题参数和考生能力的估计变得复杂（Andersen，1973；1977；Barndorff-Nielsen，1978；Rasch，1968）。对 Rasch 模型适用性的争议还涉及一个测量哲学问题：究竟应该让数据拟合模型，还是让模型拟合数据？使用 Rasch 模型的动机是让数据拟合模型，非模型拟合数据（Yu & Popp Osborn，2005）。虽然完全符合模型的试题是不存在的，但至少可以找出那些与模型偏离较远的试题。Rasch 模型下有两个检查试题拟合模型情况的指标：加权的均方拟合统计量（Infit MnSq）和未加权的均方拟合统计量（Outfit MnSq）。当试题区分度较低时，加权的均方拟合统计量往往表现为非拟合（outfit）；当试题蕴含明显猜测因素时，未加权的均方拟合统计量容易表现为非拟合（Wright，1995）。所以，与 Rasch 模型偏离较大的试题一般为区分度较低或猜测度较高的试题。

Rasch 模型对试题高区分度和最小猜测度的要求与语言测试领域评估试题质量的标准是一致的：语言测试理论以信效度衡量试题质量，试题的区分度和猜测度与试题的信效度密切相关，蕴含明显猜测因素的试题一般不具有效度。试题区分度决定试卷信度（Bachman，2004；Crocker & Algina，2008；李筱菊，2001）。事实上，Rasch 模型与数据的拟合程度提供关于考试信效度方面的信息（Linacre，2000），是试题开发和质量分析的有用工具（Clark，2007；DOLLEY，2010；刘建达，2007）。虽然 Rasch 模型拟合试题数据的程度不如二参、三参模型（Choi & Bachman，1992），但通过删除或修改拟合程度较低的试题，可以使试题数据尽量符合模型。可见，使用 Rasch 模型是在权衡测试准确性和可行性时一个折中的解决办法（Szabó，2008）。

除了要求试题具有较高区分度和较低猜测度，Rasch 模型还假设试题之间相互独立（Baghaei，2008）。这不太符合英语试题的结构特征：英语试题常常设置在同一考试材料之下，同一材料下的试题存在相互提示作用（Lee，2004）。研究表明，使用不考虑试题相互作用的 IRT 模型会高估试卷信度

(Wang et al.,2002)。Wang 和 Wilson 提议使用 Rasch 题组模型校准英语客观试题(2005)。Rasch 题组模型在 Rasch 模型的基础上引入题组反应理论(Testlet Response Theory)(Wainer et al.,2007)的题组随机参数。Rasch 题组模型不但给出试题难度值,还估计出试题间相互提示作用大小。在 Rasch 题组模型下,正确回答试题概率除了由应试者能力和试题难度决定外,还受一题组随机参数 $\gamma_{id(j)}$ 影响:

$$P_j(\theta_i) = \frac{\exp(\theta_i - b_j - \gamma_{id(j)})}{1 + \exp(\theta_i - b_j - \gamma_{id(j)})} \tag{1.2}$$

其中,$\gamma_{id(j)}$ 表示考生 i 在题组 d(j)上的题组随机参数。试题间的相互提示程度由题组随机参数的方差反映。现有文献大多讨论题组模型的适用性(Wainer et al.,2000;Wainer et al.,2007;Wainer & Kiely,1987;Wainer & Wang,2001;Wang & Wilson,2005;Wang et al.,2002),未对题组效应成因作进一步研究,而且大多使用二参或三参题组模型(Paap & Veldkamp,2012;Wainer et al.,2007;Wainer & Wang,2001),探讨 Rasch 题组模型校准英语客观试题可行性的研究还比较少。

不同试卷的试题难度值不在同一量尺上,必须先将多套试卷衔接(linking)起来,才能进一步调整各试卷应试者的原始分。现有文献等值试卷的数量一般为两套或者三套,一般不超过五套(Cook & Eignor,1991;Luo et al.,2001;Mohandas,2005;Morgan,1982)。国内外文献进行试卷等值时大多使用不考虑试题相互作用的统计模型(如 Rasch 模型(Morgan,1982;桂诗春,2000b;朱正才等,2003),二参模型(Saida & Hattori,2008;朱正才,2005))。使用题组模型等值多套英语平行试卷的文献还不多。

本文讨论 Rasch 题组模型校准和等值英语客观试题的可行性,具体回答以下三个问题:

第一,Rasch 题组模型的参数估计结果与二参、三参题组模型的参数估计结果是否一致?

第二,使用 Rasch 题组模型分析试题相互提示作用成因的结果与使用二参、三参题组模型的结果是否一致?

第三,与二参题组模型相比,使用 Rasch 题组模型等值多套英语平行试卷

时能否校正试卷间的难度差异?

本书分为三章。第一章介绍试卷等值的概念、原理方法,及试题质量各种指标;第二章介绍等值实例,对等值结果作讨论;第三章进一步讨论大规模标准化英语试卷等值的若干实际问题,并指出今后的研究方向。

第一章　试卷等值原理介绍

试题校准和等值是试题库建设的重要环节。本章首先介绍试卷等值的概念。试卷等值有两个特点：一是必须基于某种数学模型；二是必须在一定的等值条件下进行。本章介绍在项目反应理论（Item Response Theory，IRT）下如何进行试卷等值，以及满足试卷等值需求的考试设计。由于本书探讨使用Rasch题组模型进行试卷等值的可行性，本章还介绍Rasch统计模型（Rasch模型和Rasch题组模型）的特点以及用于分析试题质量的各种指标。最后，本章介绍Rasch统计模型的参数估计原理，并提供一个进行Rasch题组模型参数估计的自编程序的正确性检验结果。

1.1　试卷等值

1.1.1　试卷等值的概念

不同试卷下的分数差异主要由两个因素共同决定。一是应试者本身的能力差异，二是不同试卷试题的难度差异。试卷间的试题难度差异导致不同试卷下的分数无法直接比较。试卷等值是举行标准化考试的一项基本工作（桂诗春，2000a）。具体说来，试卷等值主要满足标准化考试两种不同的需求。试卷等值使参加不同年份考试应试者的成绩可以相互比较。等值不同年份试卷除了满足考试的公平原则外（朱正才，2005；朱正才等，2003），还可以让教育工作者和有关研究人员观察考试能力水平的变化趋势，为教育改革提供参考（Saida & Hattori，2008；桂诗春，2000c）。实行“一年多考”考试模式时，为了达到人才选拔的目的，需要使参加不同场次考试应试者的成绩可以相互比较。

第二种需求下的试卷等值往往需要同时等值多套平行试卷。

所谓等值，就是根据试卷间的难度差异调整应试者在那些具有类似内容和形式的试卷下的成绩，即使考生参加了不同的考试，试卷内容和难度不同，考试成绩也可以相互比较（Kolen & Brennan，2004）。试卷等值的目的不是消除试卷间的难度差异。当然，如果能消除试卷间的难度差异，试卷等值则无需进行。然而，产出两套难度完全相同的平行试卷几乎是不可能的。

1.1.2　试卷等值的特点

试卷等值有两个特点：一是必须基于某个数学模型或某种统计理论；二是必须在特定的条件下进行。古典测试理论中（Classical Test Theory，CTT），试卷等值主要通过线性等值和百分位数等值来调整考生的原始分。使用古典测试理论无法等值试题参数值，不能满足试题库建设的需求。在项目反应理论下，试题特征参数值和考生能力值放在同一量尺上，二者相互独立（Rentz & Bashaw，1975）。就此而论，并不需要进行试卷等值。但零点设置是确立量尺的前提条件。例如在建立测量温度的量尺时，将水的冰点确定为摄氏零度。在项目反应理论下确定洛基量尺时，同样需要确定零点。确定零点比较常见的做法是选择一基准试卷（base form）作为零点，然后将其它试卷的参数值转换到基准试卷的量尺上，通常称后面的参数值转换过程为衔接（linking）。与古典测试理论下的试卷等值相比，基于项目反应理论的试卷等值计算简洁（丁树良等，2003），优越于古典测试理论的试卷等值（Hambleton et al.，1991）。可以说，试卷等值使项目反应理论发挥出最大优势（Zimowski et al.，2003）。

试卷衔接不只是单纯的数学运算，还需要具备一定的条件，那就是两套被衔接的试卷间必须有共同考生（common examinees）或共同试题/锚题（common items）作为衔接的基础。这两类衔接条件可以大致在三种施测方法下得到满足。第一种方法是从目标考生群体中选出两组考生。这两组考生都是从目标群体中随机抽出组合而成的。可以认为两组考生的能力水平分布大致相同，这两组考生被看为是等价组（equivalent groups）。安排两组考生分别参与到两套试卷的考试中。此施测方法称为随机组设计（random group design）（图 1.2）。第二种方法是直接让同一批考生参与到两套试卷的考试

中。此施测方法称为单一组设计(single group design)(图 1.1)。这两种设计在实际考试中的可行性较低(Kolen & Brennan,2004)。采用随机组设计时,很难保证两个随机组完全等价。单一组设计受施测顺序(order effects)影响较大(Hambleton et al.,1991)。由于增加了一倍的试题量,被试回答第二套试卷的问题时容易变得疲惫,不耐烦。如果让被试在不同时间段参加两套试卷的考试,也很难保证被试具有同样的精神状态。

以上两种方法都使用共同考生作为试卷衔接的基础。第三种方法通过锚题将两套试卷衔接起来。这种方法称为非等组设计(nonequivalent groups design)(Kolen & Brennan,2004),又称为 NEAT(nonequivalent groups with anchor test)设计(von Davier et al.,2004)(图 1.3)。参与两套试卷考试的两组考生能力水平可以有一定的差异,但两套试卷间有部分试题(锚题)是完全相同的。NEAT 设计具有较高的可行性,是试题库建设常用的施测方法(Lord,1980;Szabó,2008;Vale,1986)。

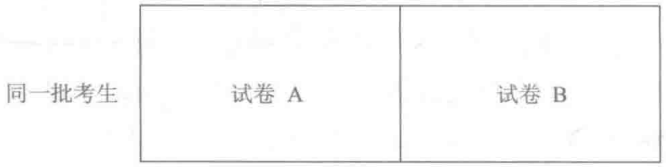

图 1.1 单一组设计

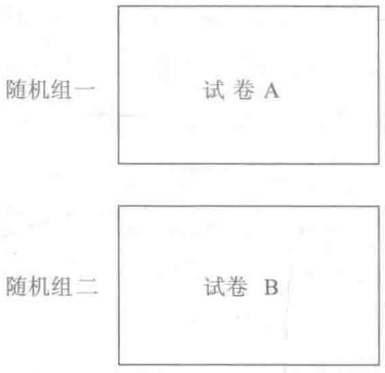

图 1.2 随机组设计