Admission & Examination Research

招生考试研究

2008

2

总第6期

上海市教育考试院主办

上海教育出版社

图书在版编目（CIP）数据

招生考试研究.6/上海市教育考试院主编. —上海：上海教育出版社，2008.9
ISBN 978-7-5444-1772-3

Ⅰ.招… Ⅱ.上… Ⅲ.入学考试—研究—中国—文集
Ⅳ.G427.74-53

中国版本图书馆CIP数据核字(2008)第167236号

招生考试研究（6）
上海市教育考试院主编
上海世纪出版股份有限公司
上 海 教 育 出 版 社 出版发行
易文网：www.ewen.cc
（上海永福路123号 邮政编码：200031）
各地新华书店经销 昆山市亭林印刷有限责任公司印刷
开本 787×1092 1/16 印张6.5
2008年9月第1版 2008年9月第1次印刷
印数 1—2,000本
ISBN 978-7-5444-1772-3/G·1433 定价：25.00元
（如发生质量问题，读者可向工厂调换）

招生考试研究
2008.2

目 录

上海市教育考试院
主　办

主　编
李瑞阳

执行编辑
谢大均
李立峰

通讯地址
上海市钦州南路500号
《招生考试研究》

邮编
200235

电话
021-64946609
021-64946607

传真
021-64513402

E-mail
zkyj@shmeea.edu.cn
xdllf2000@163.com

Contents

特　稿

用等值映射技术来检验州考标准的跨时段不变性

[美]钱家和　著　王钢　译　钱家和　校

[内容摘要]　本文研究探讨使用等值映射技术来检验各州考试能力水平标准的跨时段不变性。首先,本文提出的方法把州考能力的熟练标准映射到 NAEP(The National Assessment of Educational Progress,美国国家教育进展评估)的量表上。然后,并非直接检查州考熟练标准是否偏离原标准,而是检验各州州考标准的 NAEP 等价值的跨时段不变性。本方法的基础是一种改进的等值映射技术。在各州为公立学校学生所定考试本身相当的情形下,该技术最初用来比较各州所制定的能力标准的差异。这一检验州考标准的跨时段不变性的方法也可用来检验州考是否有分数贬值。

[关 键 词]　熟练标准;能力标准偏离(DPS);分数贬值;等百分位链接;NAEP 等价值;NCLB

1. 引　　言

当前,各州学生考试的能力标准的稳定性是美国教育界的一个关注点。根据《不让一个孩子掉队法案》(No Child Left Behind Act,NCLB),各州可以选择本州的考试并制定本州的阅读与数学的各个能力标准,以确定各州学生在规定的时间内取得长足的进步(American Federation of Teachers, 2006)。本研究旨在开发一种方法,检验州考或类似考试能力标准跨时段的不变性。能力标准表示学生能力达到熟练掌握所教的知识和技能的各种水平,通常用考试量表上的临界分数来锚定;临界分数通常把学生的能力区分为如下几个范畴:基础、熟练、高级。

如果州考的过程没有发生显著的变化,并采用等值过程来保持量表不变,熟练标准则保持在当初制定的临界分数不变。但是随着时间变迁,熟练标准会偏离量表所建立的能力水平。这种现象称为能力标准偏离(deviation in proficiency standards,DPS)或能力标准差。许多研究者发现,在与其他稳定的测试,如美国国家教育进展评估,进行比较时,一些能力评估考试存在着 DPS(NAEP;Cannell, 1987;Grissmer, Flanagan, Kawata, & Williamson, 2000;Klein, Hamilton, MaCaffrey, & Stecher, 2000;Neill et al., 1997;Smith, 1991)。另一个与考试的量表稳定性有关的概念是量表漂移。正如 Angoff(1984,p. viii)所指出的,这种漂移的产生通常由于,在将新版本的考试的量表等值转换为以往版本考试的参考量表(即作报告

作者简介:钱家和,男,美国教育考试中心(ETS)资深研究员。

译者简介:王钢,男,华东师范大学公共管理学院教授。

用的量表)时,等值转换过程不够精确所致。

很多因素会导致 DPS,例如分数贬值、量表漂移、考试工具的变化、测试形式的变化、学科框架的变革、内容修订、及差别性表现得益(Koretz, 2007; Madaus, 1988b)。然而,如果其他因素不存在,DPS 则可作为衡量分数贬值的一个统计指标。这里分数贬值指,在每个给定学术能力水平上,学生得到比以往考试更高的分数(Arenson, 2004; Koretz, 1988; Linn, 2000; Potter, 1979)。作为检验 DPS 的一种重要应用,第 4 部分将具体讨论检验分数贬值的程序。

检验熟练标准不变性的方法是基于一种改进的等值映射方法。在州考可比的情形下,该方法最初用于比较各州所制定的公立学校学生能力标准(Braun & Qian, 2007a)。很明显,只靠观察一个考试自身的分数变化,很难检验熟练标准是否偏离其初始的量表;但一个可能有 DPS 的考试与一个没有 DPS 的考试进行比较,则可检验熟练标准是否会有偏离。这方面,许多教育工作者已对州考与 NAEP 的评估进行比较。他们发现用于高利害决策的州考,其考试分数的提高往往并不能被 NAEP 分数的变化所证实(Haney, 2002; Linn, Graue, & Sanders, 1990)。这一基本策略被用来检验州考标准的不变性,即把在州考量表上的标准映射到 NAEP 量表上,以 NAEP 量表为基准再来进行比较。

实施此方法,首先要把州考的各个标准映射到 NAEP 量表上。州考标准的射像称为州考标准的 NAEP 等价值或 NAEP 等价值。虽然并没有直接探查州考标准的不变性,本文在 NAEP 量表上来探讨州考标准射像的不变性,即 NAEP 等价值的跨时段不变性。NAEP 能作为比较的基准,是由于 NAEP 在考试设计、考试内容及心理计量质量等方面,被公认为达到了高标准。因此,映射能使得跨时段比较变得有效。NAEP 是在各州施测的唯一的全国标准化考试,它统一而稳定。此外,NAEP 的分数也没有受到分数贬值等因素的影响。有关 NAEP 的简介,可参见 Jones & Olkin 的文章(2004)。尽管文献表明,由于各种原因,在学生层面上,把州考与 NAEP 测试链接起来并不是一种适当或有效的链接(Feuer, Holland, Green, Bertenthal, & Hemphill, 1998; Koretz, Bertenthal, & Green, 1999);但最近许多研究表明,对能力标准,如把州考的熟练标准,映射为 NAEP 等价值是有效的(Braun & Qian, 2007b; McLaughlin & Bandeira de Mello, 2003)。这些研究发现,大多数州考的 NAEP 等价值之间存在着很大差异,而且它们与 NAEP 本身标准之间也都不同。得到的结论是,这种异质性可归因于各州原来制定标准的严格程度的差异。

本研究假定,对同一学科,州考与 NAEP 这两种考试都是适当而对等的。又假定,州考都经过等值变换而且其各种能力标准的临界分数点不随时间而变动。此外还假定,州考与 NAEP 两种考试的其他各种考试条件,包括测试工具,也均随时间保持基本不变。

当检验发现跨时段 NAEP 等价值之间具有显著不同,这表明州考的能力标准已偏离其初始量表。但是对所观察到的变化的原因要作出判断,本文建议要成立由考试专家和学科专家组成的委员会,来确认导致显著 DPS 的原因。在对分数贬值下断言时,尤其要听取专家委员会的意见。

本文的第 2 部分将描述映射州考能力标准的估计方法,并介绍映射能力标准的跨时段特性。第 3 部分介绍研究所使用的数据,即 2003 与 2005 年四年级与八年级学生的阅读和数学州考,并展示检验州考标准不变性的实证结果。第 4 部分利用提出的方法探讨州考的分数贬值。第 5 部分是总结与一些结论。

2. 方　　法

本章先介绍把州考能力标准映射到 NAEP 量表上的过程，然后叙述检验州考能力标准不变性的方法。

2.1　把州考标准映射到 NAEP 量表上的方法概述

这里介绍的映射过程是由 Braun & Qian(2007a)提出的。在分析中，利用全国纵向学校层面州测试分数数据库①，对参加 NAEP 测试的各州的数据都分别实施了这个映射过程。要能有效地比较跨时段的 NAEP 等价值，州考及 NAEP 都必须符合一定的条件。在 2.2 节将介绍其必须符合的标准条件。本研究的统计分析涉及 NAEP 抽样设计、学校权重、标的估计等事项。NAEP 采用两阶段概率抽样设计获取各个州的样本。考虑到选取概率的不均等以及要对未答进行调整，每所学校及每名学生分别赋予抽样的权重。本研究在估计州达标学生的比例时采用了适当的权重，并且利用了比率估计量。本文的附录 A 提供了对权重、标的估计以及方差估计的简要描述。

设 P 表示达到一种特定标准学生的全州比例。设 F 表示该州在 NAEP 测试中的分数分布，F 的第 $(1-P)$ 分位数为 $\xi = F^{-1}(1-P)$。第 $(1-P)$ 分位数的估计值 $\hat{\xi}$，可以用 $\hat{\xi}_{WAM}$ 表示，其中缩略词 *WAM* 代表“加权总计映射”。把州考标准映射到 NAEP 量表上的过程，Braun 和 Qian(2007a)设计按下列步骤进行：

1. 依据本州 NAEP 样本中各校在州考中达标的学生比例，来估计该州达到州考标准的全州学生比例。

首先，找出在州 NAEP 样本中的学校，并与其在全国纵向学校层面州测试分数数据库(NLSLSASD, the National Longitudinal School-Level State Assessment Score Database)的记录相匹配。得到每所学校达到州考标准的学生比例。采用 NAEP 设计的学校权重，用比率估计值 $\bar{p}_w$(该值为达到标准学生数的加权平均估计值与有资格参试学生数的加权平均估计值之比)，可以得到 p 的估计值。更详细的权重及比率估计值描述，可参见附录 A。

2. 依据 NAEP 学校与其校内学生的样本，估计全州在 NAEP 测试中的分数分布。这一步产生的结果，也会刊登在每次 NAEP 测试后所发表的报告中。设 F 表示 F 的经验分布，可依据 NAEP 样本获得。

3. 在 NAEP 量表上找到一点，在该点上使达标的学生比例等于州考达标的学生比例。

据上面按步骤 1，用 $\bar{p}_w$ 来估计州考达标的学生比例 p（其界定在州考量表上），及按步骤 2 计算出 NAEP 分数分布之后，在 NAEP 量表上找出 $\hat{\xi}$ 点，即第 $(1-\bar{p}_w)$ 分位数，把表现标准映射到 NAEP 量表上：

$$\hat{\xi}_{WAM} = \hat{F}^{-1}(1-\bar{p}_w) \tag{1}$$

---

① 全国纵向学校层面州测试分数数据库(NLSLSASD；www. school. org)是由美国研究所(AIR)为全国教育统计中心(NCES)构建并维护的。其宗旨是收集与验证全国州级测试项目的数据。它存有美国近 80000 所公立学校的测试数据，并每年更新。

被估计的州考标准的 NAEP 等价值是 ξ 的估计值，可用 $\hat{\xi}_{WAM}$ 表示。如果州的标准不止一个，对每个标准都可用这一过程来估计。

4. 估计被估计的 NAEP 等价值的方差。

这里的计算是根据 NAEP 折刀方法（jackknife methods：一种用于减少偏差及估计方差的统计方法，译者注）发展而来。依据给定的 NAEP 抽样设计，用折刀方法对潜在能力分数来估计其方差（Allen，Donoghue，& Schoeps，2001）。

图 1 阐释了映射过程。左边的间断曲线代表根据州 NAEP 样本学校中全体州考学生的分数，得到的州考分数的分布曲线。该分布上端高于州考标准的面积为该州达到或超过标准的学生比例的估计值，用 $\hat{p}_w$ 表示。在实践中，$\hat{p}_w$ 是仅需从数据中求得的量。右边的曲线代表该州 NAEP 分数的分布曲线。这是通常报告的 NAEP 分布，其可根据参加 NAEP 测试的州的 NAEP 学生样本来估计。NAEP 对州考标准的等价值 $\hat{\xi}$，是 NAEP 量表上的一点，使 NAEP 分布的相应上端面积也等于 $\hat{p}_w$。假定州考分数分布以及 NAEP 测试分数分布不变，由于等百分位链接的单调特性，较大的 $\hat{p}_w$ 对应于较小的 $\hat{\xi}$，反之亦然。

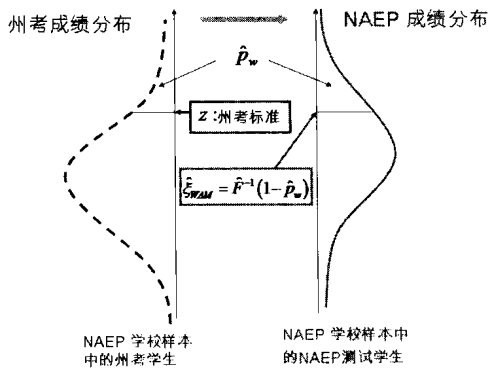

图 1　映射过程的示意图

2.2　检验跨时段州考标准的不变性

把州考标准映射到跨时段的 NAEP 量表上。如引言所述，对州考与 NAEP 两个测试间作等值映射，映射方法的有效性要求两个测试适度的对等，包括学科框架、测试版式、考试的心理计量特性、常模等方面的对等。这个过程也要求这两个测试符合下列三个条件：（a）考试随时间没有较大的变化；（b）考试要施行等值变换，并且划定标准的临界分数不随时间变化；（c）考试分数的分布曲线在不同时段保持相同的形状和宽度，但允许分布曲线有水平的漂移。上述标准条件似乎很严格，然而对一般考试而言却都是合理的努力方向。

设 z_A 与 z_B 分别表示时间点 A 与 B 的州考标准。由于假定州考标准的临界分数不随时间变化，因此，$z_A = z_B$。设 ξ^A 与 ξ^B 分别表示时间点 A 与 B 的 z_A 与 z_B 的映像。它们的估计值为 $\hat{\xi}^A$ 与 $\hat{\xi}^B$。$\hat{\xi}^A$ 与 $\hat{\xi}^B$ 的方差估计与 2.1 节中 $\hat{\xi}$ 的方差估计相同。

设 P^A 为时间点 A 达到标准 z_A 的学生比例，而 P^B 为时间点 B 达标的学生比例。图 2 左

边的两条经验曲线阐释了两个时间点之间的变化，而图右边的 $\hat{\xi}^A$ 与 $\hat{\xi}^B$ 则是映射过程的结果。

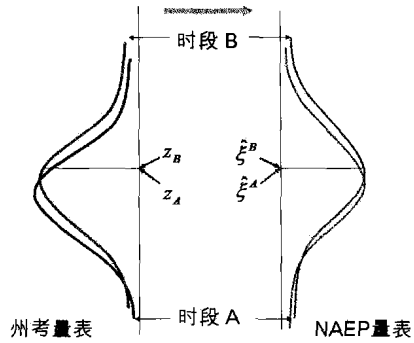

图 2　两个时段州考的映射过程

设 $P^B = P^A + \Delta P^S$，其中 ΔP^S 表示在州考中达标的学生比例的变动。当 $\Delta P^S > 0$ 时，表示在时间点 B 上达到标准的学生比例更高。在时间点 B 上达到标准的学生比例更高可能源于如下一个或两个原因：教育取得了真实的进步，或者考试结果存在 DPS。如果是教育取得进步，可以假定学生应当在州考以及相应的 NAEP 测试中显示相似程度的进步。

NAEP 有跨时段等价值的一些特性。设 F^A 与 F^B 分别表示时间点 A 与 B 的 NAEP 量表估计分布。如在(1)中所示，时间点 A 的 NAEP 等价值 P^A 的映像，是 F^A 的第 $(1 - P^A)$ 分位数：

$$\xi^A = F^{-1,A}(1 - P^A) \tag{2}$$

而 P^B 对 F^B 的映像是

$$\xi^B = F^{-1,B}(1 - P^B) = F^{-1,B}(1 - (P^A + \Delta P^S)) \tag{3}$$

设 P^α 表示在时间点 B 在 NAEP 本身测试中得分高于 ξ^A 点的学生比例，即

$$\xi^A = F^{-1,B}(1 - P^\alpha) \tag{4}$$

由于学生跨时段表现的不同，P^α 通常不等于 P^A。在时间点 B 令 $P^\alpha = P^A + \Delta P^N$，因而 ΔP^N 是在 NAEP 量表上高于 ξ^A 点的比例的变动。

首先，假定 $\Delta P^S = \Delta P^N$，即该时段学生在州考达标比例以及相应的 NAEP 测试达标比例中显示出相同的变动。这意味着 $P^\alpha = P^A + \Delta P^S$。鉴于(4)和

$$\xi^B = F^{-1,B}(1 - (P^A + \Delta P^S)) = F^{-1,B}(1 - P^\alpha) \tag{5}$$

因此 $\xi^B = \xi^A$。这一结果表明当 $\Delta P^S = \Delta P^N$ 时，跨时段的 NAEP 等价值是不变的。相应地，ξ^A 可被视为一个跨时段不变等价值。图 2 阐释了所讨论时段州考以及 NAEP 测试两者表现的映射过程。在标准条件下，用 NAEP 量表作为比较的基准，跨时段 NAEP 等价值的不变性等价于跨时段州考能力标准的不变性。

其次，假定 $\Delta P^S > \Delta P^N$，即 $P^A + \Delta P^S > P^A + \Delta P^N$，在州考中达到标准的学生比例变动高于 NAEP 测试的相应比例变动。由于

$$\xi^A = F^{-1,B}(1 - P^\alpha) = F^{-1,B}(1 - (P^A + \Delta P^N)) \tag{6}$$

以及 $F^{-1,B}(g)$ 的单调性，可得出

$$\xi^B = F^{-1,B}(1-(P^A+\Delta P^S)) < F^{-1,B}(1-(P^A+\Delta P^N)) \tag{7}$$

即 $\xi^B < \xi^A$，这表明时间点 B 的 NAEP 等价值比 ξ^A 低。它显示了 DPS 的出现，既对能力标准的偏离。图 3 上的经验性映射过程，阐释了跨时段的州考表现与 NAEP 测试之间的不同。

第三，假定 $\Delta P^S < \Delta P^N$，即 $P^A + \Delta P^S < P^A + \Delta P^N$，在州考中达到标准的学生比例变动小于 NAEP 测试的相应比例变动。由公式(6)以及 $F^{-1,B}(g)$ 的单调性，可得

$$\xi^B = F^{-1,B}(1-(P^A+\Delta P^S)) > F^{-1,B}(1-(P^A+\Delta P^N)) \tag{8}$$

即 $\xi^B > \xi^A$。这种假定的况态并不常现，尽管它显示了 DPS 的出现。

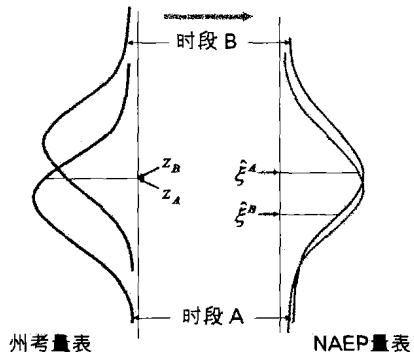

图 3　在时间点 B 州考分数贬值的映射过程

检验 NAEP 等价值跨时段不变性。在本研究中，评价程序同时采用统计显著性检验以及效应量准则。统计检验的假设用来核查在标准条件下跨时段 NAEP 等价值的不变性。零假设可表达为 $H_0: \xi^B = \xi^A$。此假设等价于在时间点 B，高于 ξ^B 的学生比例是否等于高于不变等价值的比例：$P^B = P^\alpha$。分析中采用了两种显著性检验。第一种检验用 t - 类型统计量，核查两个比例的差异。第二个统计量是 log-odds 比率(Haberman，1978)。

设 n_B 为时间点 B 所考虑的样本容量。设 $\hat{\xi}^B$ 与 $\hat{\xi}^A$ 分别为 ξ^B 与 ξ^A 的估计值。在表 1 中，设 n_{11} 与 n_{21} 分别为得分高于 $\hat{\xi}^B$ 与 $\hat{\xi}^A$ 的学生数，而 n_{12} 与 n_{22} 分别为未达到标准的学生数。设 $\hat{p}_w^B = n_{11}/n_B$ 为 P^B 的估计值，而 $\hat{p}_w^\alpha = n_{21}/n_B$ 为 P^α 的估计值。设 $\hat{p} = n_1/n$，而 $q = 1-p$。定义 Z_c 统计量为

$$Z_c = \frac{|p_w^\alpha - p_w^B| - 1/n_B}{\sqrt{2}pq/n_B} \tag{9}$$

式(9)中的 $1/n_B$ 项是连续性 Yates 校正(Yates，1934)。log - odds 比率定义为

$$L = \log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right) \tag{10}$$

以及其标准误差的估计值为

$$SE(L) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \tag{11}$$

由于 NAEP 采用两阶段抽样方法收集州数据，简单随机抽样公式会低估检验统计量的

方差。复杂数据的方差估计通常使用重复再抽样方法(Wolter,1985)。为了简化计算并考虑复杂抽样的效应,采用方差估计值与设计效应的乘积来估计方差,统计量设计效应是由Kish(1965)引进,其等于复杂抽样统计量方差与简单随机抽样统计量方差的比率。根据以往的 NAEP 分析,在计算时采用 2.5 作为近似的设计效应估计。然后在统计检验的分析中采用 0.05 的 α 水平。

表 1　在时间点 B,得分高于 $\hat{\xi}^B$ 及通过 $\hat{\xi}^A$ 的 *NAEP* 学生数

	能力标准		
	通过	未通过	总计
$\hat{\xi}^B$	n_{11}	n_{12}	n_B
$\hat{\xi}^A$	n_{21}	n_{22}	n_B
总计	n_1	n_2	N

当假设被拒绝,这表明跨时段的 NAEP 等价值有显著差异,意味着具有明显的 DPS。它表明学生在州考表现不同于其在 NAEP 测试中应有的表现。然而,DPS 并不能视为等于分数贬值,因为其他可能因素,包括差别性表现得益和课堂教学风格,也会导致这些差异。只有当其他潜在因素能被排除时,DPS 才能作为分数贬值的一种指示物。

实际运用上,效应量准则(effect size)也用于评价来自独立样本的两种特性的差异,或单一比例与任何特定假设值之间的差异。比例比较的效应量称为 H 指数。为了让差异在整个量表上,更好地使比例的效应量变化均匀,较容易察觉变动,Cohen(1988)提出在计算差异前,先对比例进行反正弦变换。设反正弦变换 $\varphi = 2\arcsin\sqrt{p}$。比例的 H 指数定义为 $H = |\varphi_1 - \varphi_2|$。考虑到其作为一个中间的效应量,在测量两个比例的差异时,H 指数的绝对值至少应当为 0.20。

2.3　评价考试结果

在检验出明显的 DPS 后,发现导致偏离标准的原因也很重要。要测试检验结果,应成立由考试专家及学科专家组成的委员会。这一过程类似于 NAEP 的 DIF (differential item functioning)分析的审查过程(Allen et al., 2001)。

为了判断偏离跨时段 NAEP 等价值的原因,整个过程由两阶段组成。第一阶段涉及进行相关的计算以及统计检验。第二阶段涉及测试结果以及确认可能导致跨时段能力标准偏离的因素。专家委员会将核查标准条件的假定,审查发现的结果,并讨论引起差异的可能原因,以及得出结论。只有排除了所有的潜在原因之后,结果才能将 DPS 归因于分数贬值。

3. 经验数据的应用

3.1　数据

为了探查跨时段熟练标准的偏离,本研究分析了两组数据:(a)2003 与 2005 年 NAEP 数学与阅读测试的四年级和八年级学生样本,(b)2003 与 2005 年州数学与阅读考试的四年

级和八年级学生样本。达到2003与2005年州考标准的学生特性信息从全国纵向学校层面州测试分数数据库(NLSLSASD)中获取。该数据库始建于1994学年,存有几乎所有州,按学校分类的,达到州各类标准的学生的比例信息。然而,它并没有学生个体的分数。全国纵向学校层面州测试分数数据库通常呈现每所学校达到并高于州所制定每一能力标准的学生百分比。①

3.2 经验性结果

首先完成2.1节所描述的映射过程。在Braun和Qian所写的报告中(2007b),有所计算的估计值的表格,其中包括2005年四年级、八年级州阅读与数学考试的全州考熟练学生比例、州考标准的NAEP等价值、州考标准的NAEP等价值的标准误差。还包括2003年四年级、八年级州阅读与数学考试的同类结果。每个表列出现各州NAEP样本学校数以及映射使用的学校数。最后的这个数值只是能与有可用的州考表现数据学校相配对的NAEP样本学校数。各表下面还注释列出了数据分析中的有关问题。为了显示等值映射方法在比较各州标准中的应用,附录B包括了有关2005年四年级、八年级NAEP阅读和数学测试的州考标准的NAEP等价值结果的四张图。由图B1可见四年级NAEP阅读测试的NAEP等价值最高和最低之间相差达75分之多,表明各州州考标准设定互相之间相差很大。图B2~B4也显示了相同的情况。各州间考试标准设定的差异给评估各州教育进步带来了困难,因而成为公平执行NCLB政策的一个难题(Lewin, 2007)。

在四年级阅读分析中,比较了同时具有2005年与2003年数据的25个州中21个州的数据。为了使州考与NAEP阅读测试尽可能匹配,如果州测试的名称为"英语/语言艺术"而不是"阅读",该州的数据就不予采用。此外,只讨论达到标准学生比例增加的那些州。分析结果显示有两个州在统计检验与效应量两方面核查均具有显著性的结果。这两个州列在表2中(州代号为1和2)。注意:因为尚未对偏离熟练标准的可能原因进行调查,表2中仅列出州代号而非州名。分析检验显示,代号为1的州达到其标准的学生比例有较大的提高。在其2005年NAEP样本中,通过ξ^B的学生比例是0.71,而通过ξ^A的学生比例是0.60。在$\hat{F}^A$上的$\bar{p}_w^A$映像(.60)以及在$\hat{F}^B$上的$\bar{p}_w^B$映像(.71)显示在跨时段NAEP量表具有显著的变异。这表明出现了明显的DPS,即州考熟练标准有了偏离。在八年级阅读分析中,比较了同时具有2005年与2003年数据的30个州中28个州的数据。分析结果显示有五个州(州代号为3-7)在统计检验与效应量两方面核查均具有显著性的结果。这五个州的结果列在表2中。

在四年级数学分析中,比较了同时具有2005年与2003年数据的25个州中24个州的数据。第一阶段分析后,列在表2中的三个州(州代号为8-10)显示NAEP等价值在统计检验与效应量两方面核查均具有显著差异。其中,代号为8的州考试显示达到其标准的学生比例显著增加。在2003年与2005年分别有74%和85%的学生达到标准。在其2005年的NAEP样本中,达到其标准ξ^A的学生比例为0.76。检验显示,在$\hat{F}^A$上的$\bar{P}_W^A$映像(.71)以及在$\hat{F}^B$上

① 在几乎所有州,NAEP学校样本中的一些学校数据,或未收入NLSLSASD,或所需的数据并未列出。在这些场合,用于估计的学校数就小于NAEP学校样本中的学校数。对于每一学科与年级的组合,有四至五个辖区,其用于估计的NAEP学校样本比例不到0.9。

的 $\overline{P}_W^B$ 映像(.85)的变异是显著的。它意味着州代号为 8 的四年级数学考试具有明显的 DPS，或州考熟练标准发生了偏离。要查证能力水平百分比变化的原因，必须进行进一步的调查，并在第二阶段中获得专家委员会的最终认可。在八年级数学分析中，比较了同时具有 2005 年与 2003 年数据的 32 个州中 25 个州的数据。分析结果显示有五个州(州代号为 11 - 15)统计检验与效应量两方面核查均具有显著性的结果。

表 2　*NAEP* 阅读和数学测试统计检验和 *H* 指数核查具有显著性的结果

州代号	2005：超过 $\hat{\xi}^B$ 比例的估计值 $\hat{\overline{p}}_w^B$	2005：*NAEP* 等价值 $\hat{\xi}^B$ 的估计值	2005：超过 $\hat{\xi}^A$ 比例的估计值 $\hat{\overline{p}}_w^\alpha$	2003：*NAEP* 等价值 $\hat{\xi}^A$ 的估计值	Z_c 统计量	Log-odds 比率统计量	H 指数统计量
四年级阅读：							
1	0.71	202	0.60	212	6.61	0.21	0.23
2	0.80	197	0.67	210	6.89	0.29	0.30
八年级阅读：							
3	0.63	244	0.52	256	5.15	0.19	0.22
4	0.82	235	0.73	247	5.24	0.23	0.22
5	0.72	245	0.63	256	5.56	0.18	0.19
6	0.30	276	0.19	285	6.02	0.27	0.26
7	0.57	254	0.43	267	6.48	0.25	0.28
四年级数学：							
8	0.85	218	0.76	226	5.72	0.25	0.23
9	0.80	224	0.65	234	6.79	0.33	0.34
10	0.91	207	0.78	217	8.50	0.45	0.37
八年级数学：							
11	0.61	269	0.52	278	4.35	0.18	0.20
12	0.53	276	0.44	286	4.51	0.17	0.20
13	0.74	258	0.64	268	4.68	0.20	0.22
14	0.70	266	0.53	280	8.24	0.32	0.35
15	0.65	277	0.44	293	8.82	0.37	0.42

4. 一种应用:探查州考的分数贬值

本方法一个重要的应用是用来探查州考的分数贬值。如果能排除导致 DPS 的其他因素,那么,显著性的 DPS 就表明存在分数贬值。因此,DPS 是检验分数贬值的一个必要条件。

近年来,分数贬值之所以成为众多教育工作者关注的问题,因为它抵消了对改进教育与测试有效性的努力(Bromley, Crow, & Gibson, 1973;Hambleton et al., 1995;Rosovsky & Hartley, 2002;Shepard, 1988)。分数贬值可以与各种情形相关联。显然,一个考试没有作等值链接或其等值链接不良好,可能导致分数贬值。但是,即便一个考试具有良好的链接或等值,仍可能出现分数贬值。典型的情况是课堂教学以应付考试为动力,以及学生的学习专注的内容集中于标准化考试试题。学生能力水平虽然不同,学习均集中相同的试题与内容的,因此,结果分数并不反映个体学生的真实学术水平。尤其是,在这样的环境中,能力水平较低的学生常常获得高于其相应能力的考试分数(Haladyna, Nolan, & Hass, 1991;Madaus, 1998a;Phelps, 2005)。这些情境导致测试不能充分测量学生的能力水平,即便努力使考试与课程大纲保持紧密一致,也并不能充分防止这种类型的分数贬值(Koretz,2005)。

检验分数贬值的原理是,首先用 NAEP 分数的变化幅度来检查相应的州考分数的变化幅度,它以 NAEP 量表的稳定性作为比较的基础。然后,如果探查到了 DPS,要请专家小组判断分数贬值是否是造成 DPS 的原因。

在 2.2 节中讨论了两种 DPS 的情形,其中一个与分数贬值可能相关。当 $\Delta P^S > \Delta P^N$ 时,即意味着在州考中达标的学生比例变动大于在 NAEP 测试中的学生比例变动。同时表明时间点 B 的 NAEP 等价值低于 ξ^A;$\xi^B < \xi^A$。这一情形时显著性 DPS 提示了有可能发生分数贬值。在另一情形,即 $\Delta P^S < \Delta P^N$,其也意味着 $\xi^B > \xi^A$。这一情形时显著性 DPS 当然不是分数贬值,而表示可能未满足标准条件或测试条件有了变化。

要正式断言分数贬值,必须经过专家委员会对 DPS 的原因作出评价,以及对非分数贬值的潜在因素进行讨论后,才能正式得到。此外,也存在这样的可能性,即跨时段 NAEP 等价值的变化是由多个因素混合而形成的:部分是由于试题形式及考试结构的修改,部分是由于分数贬值。要解决此类情境的问题以及得出结论,必须在进一步的研究中收集更多的数据。

表 2 对 2005 与 2003 年的四年级与八年级的阅读和数学数据作了统计分析,尽管证明了存在显著的 DPS,人们并不能由此断定 DPS 的具体原因,包括可能的分数贬值,因为这些尚未经过专家委员会的审查。

5. 结　　论

本文为州考以及其他类似测试提供了一种检验跨时段州考熟练标准不变性的方法。该方法所依据的原理,其设计最初是用来对州际标准设定进行相互比较;在开发中,两个应用均以 NAEP 量表为基准。

该方法发展源自处理现实测试问题的需要(Thissen, 2007)。众所周知,随着时间的变

迁,分数贬值、量表漂移、差别性表现得益、考试工具结构的变化、内容修订以及课堂教学的风格都能引起分数偏离。显然,分数偏离这个概念要比标准偏离更为宽泛。表3是讨论分数偏离时展示的一个实例。它列出了肯塔基KIRIS[①]四年级阅读分数及对应的NAEP分数的变化(Koretz, 2007)。

表3 肯塔基KIRIS与NAEP四年级阅读熟练水平的变化(1992~1994)

	原始分变化	标准化变化
KIRIS	18.8	0.76
NAEP	-1.0	-0.03

显然,NAEP与KIRIS的变化大小是不同的,但进行直接比较却是不妥当的,因为它们被放在不同的量表上。尽管标准变换能校准这两个量表,但基于标准变换的检验统计量复杂,则宜用本研究所提出的方法对此类问题进行检验。

方法的全过程包括检验跨时段的DPS、对标准考试条件符合情况进行验证、由专家委员会对变化的原因进行评价。在标准条件下,NAEP等价值跨时段的实质性差异是可能的分数贬值的一种指示。总的说来,对于用这种方法报告的(差异)具有统计显著意义的任何考试,均需由考试和学科专家委员会对结果进行审查,并判定原因是否为分数贬值。只有在与考试条件相关的其他因素,如内容修订以及考试工具的变化等都可以忽略时,才能作出这一判定。对于肯塔基这一实例,要确认分数贬值是否为KIRIS考试分数变异的原因,也应当实施前面所建议的全部统计过程。

正如2.3节所提到的,NAEP等价值跨时段差异的产生原因可部分归因于考试条件的变化,部分归因于分数贬值。如存在着混合原因,将增大调查的难度。如在数据信息有限的情况下要作出任何推断,如分数贬值,都应当保持适度的谨慎。

参考文献

[1] Allen, N., Donoghue, J., & Schoeps, T. (2001). The NAEP 1998 technical report (NCES 2001-509). Washington DC: National Center for Education Statistics.

[2] American Federation of Teachers (2006). Smart testing: Let's get it right. Unpublished reviews, available from http://www.aft.org/pubs-reports/downloads/teachers/Testingbrief.pdf

[3] Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

[4] Arenson, K. W. (2004, April 18). Is it grade inflation, or are students just smarter? New York Times, p. WK2.

[5] Braun, H. I. & Qian, J. (2007a). An enhanced method for mapping state standards

① 肯塔基教学结果信息系统(KIRIS)。在1998~1999学年被联邦(教育)责任制与测试系统(CATS)所取代。

onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland, (Eds.), Linking and aligning scores and scales (pp. 313 - 338). New York: Springer-Verlag.

[6] Braun, H. I., & Qian J. (2007b). Mapping 2005 state proficiency standards onto the NAEP scales (NCES Research and Development Report No. NCES 2007 - 482). Washington DC: National Center for Education Statistics.

[7] Bromley, D. G., Crow, H. L., & Gibson, M. S. (1973). Grade inflation: Trends, causes, and implications. Phi Delta Kappan, 59(10), 694 - 697.

[8] Cannell, J. J. (1987). Nationally normed elementary achievement testing in America's public schools. How all fifty states are above the national average. Daniels, WV: Friends for Education.

[9] Cochran, W. G. (1977). Sampling techniques (3rd ed.). New York: John Wiley & Sons.

[10] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

[11] Feuer, M. J., Holland, P., Green, B. F., Bertenthal, M. W., & Hemphill, F. (Eds.). (1998). Uncommon measures: Equivalence and linkage among educational tests. Washington, DC: National Academy of Science.

[12] Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). Improving student achievement: What state NAEP test scores tell us. (Rand Corporation Rep. No. MR-924-EDU). Santa Monica, CA: Rand Corporation.

[13] Haberman, S. J. (1978). Analysis of qualitative data: Vol. 1, Introductory topics. New York: Academic Press.

[14] Haladyna, T., Nolen, S., & Haas, N. (1991). Raising standardized achievement test scores and the origins of test score pollution. Educational Researcher, 20(5), 2 - 7.

[15] Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). Review of the measurement quality of the Kentucky Instructional Results Information System, 1991—1994. Frankfort, KY: Office of Education Accountability, Kentucky General Assembly.

[16] Haney, W. (2002). Ensuring failure: How a state's achievement test may be designed to do just that. Education Week, 56, 58.

[17] Jones, L., & Olkin, I. (2004). The nation's report card: Evolution and perspectives. Bloomington, Indiana: Phi Delta Kappa International.

[18] Kish, L. (1965). Survey sampling. New York: John Wiley & Sons. Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? Education Policy Analysis Archives, 8, 49.

[19] Koretz, D. M. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? American Educator, 12(2), 8 - 15, 46 - 52.

[20] Koretz, D. M. (2005). Alignment, high stakes, and the inflation of test scores.