DOI:10.16382/j.cnki.1000-5560.2018.03.008

关于选考科目等级赋分的改进：历史经验、现实限制与可能方向

章 建 石

(教育部考试中心,北京 100084)

摘 要:由高考科目调整引发的不同科目成绩之间的可比性问题,已多次成为高考改革面临的一个难题。有关省份的改革经验表明,标准分在分数转换上有着明显弊端,并不是解决这一难题的理想方案。在原始分主导、总分录取等诸多前置条件的约束下,现行等级分也是一种简化的"标准分"。等级分在解决分数的可比性、总分的可加性上做出了有益探索,但也带来了一些不符合教学导向的消极影响。在不打破当前高考分数使用方式的前提下,等级分在考试技术层面仍具有改进的空间。

关键词:高考;分数转换;等级分

一、新改革,旧问题——科目调整与赋分方式的历史经验

目前,浙沪两地新高考改革已经完成了一个周期。从各方面反馈来看,两地试点可以说取得了显著成效,但也有一些技术细节引起了一些争论,其中包括选考科目的计分方式。对这方面的进一步完善,社会各界有很高的期望。在此,不妨先回顾最近的一段改革历史,看是否有可以吸取的经验,这也有利于认清当前的处境和改革推进的方向。

上世纪 90 年代末开始,我国高考开启了新一轮的改革历程。为主动适应时代特点及其对人才素质能力结构的要求,着力引导人才全面素质的提高和创新人才的培养,1999 年,教育部发布了《关于进一步深化普通高等学校招生考试制度改革的意见》(教学[1999]3 号),提出在高考科目设置、考试内容与形式、录取方式上进行全面的改革。该意见明确提出,在科目设置上,用三年左右的时间推行"3 + X"科目设置方案。在这一科目设置中,"3"是指每一位参加高考的考生都必须考语文、数学、外语 3 科,"X"指的是每一位考生必须从政治、物理、化学、生物、历史、地理 6 个科目或综合科目中确定一科或几科,这些选考科目由高校根据办学层次等要求来指定。其中,综合科目是指建立在中学文化科目基础上的综合能力测试。从推广的时间来看,1999 年,广东率先进行高考"3 + X"科目设置改革,2000 年,吉林、山西、江苏、浙江 4 省参加改革,2001 年扩展到 18 个省、市、自治区。2001 年 9 月,教育部批复同意北京、河北、山东、江西、广西、重庆、云南、贵州、西藏、甘肃、宁夏、新疆、青海等 13 个省、市、自治区从 2002 起进行高考"3 + X"科目设置改革。至此,"3 + X"科目设置改革在全国各省、市、自治区全面展开。

各地在落实这一改革的过程中,在科目设置上的举措也不尽相同,其中比较有特色的是广西。在广西的科目设置改革方案中,语文、数学、外语 3 门科目为考生必考科目,"X"是物理、化学、生物、政治、历史、地理和综合能力测试(以下简称"综合")以及其他技能中若干科目的组合,考生可根据自己的兴趣、爱好、特长、学习能力和选择专业的要求,自主决定选考科目。为避免科目变动过大引起不必要的争论,广西在"X"的科目设置上采取了有限放开的策略,规定当年报考本科专业要求的"X"为 2 门,可

供选择的科目组合一共有 12 种。[①]这种科目设置在当时的“3 + X”改革中,独树一帜。这次改革使学生和高校具有较大的选择权,也兼顾了统一性和选择性。统一性体现在把语文、数学、外语课程作为必考科目上,使其基础性、通用性的地位得到了认可。选择性则体现在“X”上,由高校根据专业特点来设置,再由考生自主选择。

“3 + X”改革的初衷非常明确,“进行高考科目设置改革,其中最重要的是高等学校要有自己选择考试科目的权利,学生相应地有选择应试科目的权利,这是今后改革的主要方向”(郭小川,2002)。从这点来看,广西的改革打破了考试科目设置的单一模式,在坚持统一性的前提下,比较充分地凸显出了高校与学生的双向选择,有利于高校扩大办学自主权,按各专业人才培养的需求来选拔人才。同时,对考生而言,也有利于扬长避短,充分发挥自身的学科优势,从而在一定程度上实现了减轻应考压力的目标。这样的改革立意,与 2014 年启动的新一轮高考改革,在具体举措及所倡导的选择性理念上是非常相似的。

广西方案公布以后,社会各界的反响较好,上级主管部门也给予了充分肯定,一度被认为是较为符合高校人才选拔规律的改革举措。然而,与不少改革初衷良好的方案一样——广西的这次改革最终没有经受得住实践的检验。实施第一年,由于科目组合的多样化,高校在录取过程中遇到了不少困难,顾虑重重。[②]中学的教学组织、师资配备面临巨大压力,只能在教学中减少科目组的选择。高校、中学对科目组合的态度和应对方式,限制了学生的选择范围,以致于很少有学生按照自己的兴趣和特长来进行选科。2004 年,迫于多方面的压力,原本可选择的 12 个科目组合减少了一半,剩下 6 个。到 2005 年,又迅速调整为“3 + 文科综合或理科综合”,这已与当时其他大部分省份的高考科目设置基本一致了。

至此,广西这一次颇具新意的科目组合改革如昙花一现,迅速销声匿迹。如此重大并且在设计之初看来具有种种优点的一项改革举措,还没等到一届高中生毕业就被迫终止,这在我国高考改革历史上是不多见的。其原因很复杂,难以逐一展开分析,本文仅从高考计分的角度来进行探讨,希望对当下浙沪新高考改革试点中关于选考科目赋分的改进,提供一些参照。在此,另一个背景需要交代清楚,那就是标准分制度在高考计分中的使用。这项改革从 1985 年在广东开始试点,截至 1997 年,标准分制度推广至海南、河南、陕西、广西、山东、福建等省。[③]广西推出“3 + X”的改革时,标准分制度已经在当地实施。可以说,广西的科目改革与标准分共存了一段时间,那么,在当时的改革情境中,标准分并没有解决科目设置多样化所带来的问题,相反还暴露出了标准分自身的不足。这一切导致的直接结果是:2005 年,广西在决定科目设置采用“3 + 文科综合或理科综合”时,同时取消了实施多年的标准分制度,重新使用原始分进行录取。

二、标准分是解决不同科目分数可比性的出路吗

在浙沪高考改革试点过程中,笔者多次参与国家教育咨询委员会、国家教育考试指导委员会在两地的调研、研讨、督查与总结工作。期间,对于试点方案中选考科目赋分以及高考总分的合成,一线教师、相关领域专家多次呼吁恢复曾经使用过的标准分制度。然而,上述广西的高考科目改革恰恰正是在使用标准分的同时被取消的。前车之覆,后车之鉴。改革再一次将不同科目分数的可比性难题呈现在了改革者面前。新高考改革方案中关于选考科目的赋分,标准分是不是一个好的备选方案呢?

(一)标准分的分数转换

在大规模考试中,原始分是最简单的一种分数报告方式,原因是只需统计一下答对题目的数量,参照各题的分值,经过相加就能得出分数。但是,其不足也很明显,其中很重要的一个是不具备可加性。由于不同科目试卷难度不同,各科原始分的单位和参照点也不相同,各科原始分相加得到一个总分,意义不明确,可行性也存疑。改进的办法是引入导出分数,也就是说在原始分基础上,按照一定的规则,经过统计处理后获得的具有一定参照点和单位、且可以相互比较的分数(戴海崎等,2006,第 136 页),这个过程可以称之为分数转换。分数转化所采用的方法、算法和约定的规则不尽相同,导出分数的形

态也有差异。标准分就是一种广泛应用的导出分数，它是把原始分转化为具有相同意义、相同单位和共同参照点，能表明考试成绩在总体中位置的分数。

大量数据表明，在高考中，尽管考生数量庞大，但其原始分的分布基本是偏态的，难以直接进行比较。[④]对此进行正态化处理是标准化转换的一个基础性工作。标准分转换的核心思路是：将每一个原始分分值对应的百分等级对应作为标准正态分布下的左端面积值，从而对原始分分布进行强制正态化，再按照设定的均值和标准差进行一次线性转换，从而得到在特定分数全域内的分值（见示意如图 1）。

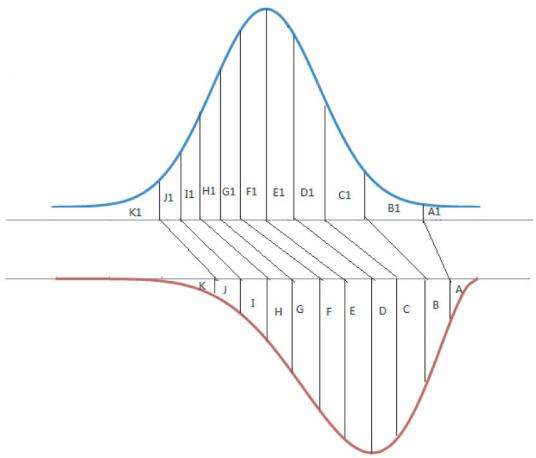

图 1　负偏态分布正态化过程

依据上述原理，我国在广东等省份推广标准分时，采取了以下几个关键的技术处理。分别是：（1）单科的正态化处理，（2）单科的标准化，（3）单科标准化后的线性转化。考生总分的合成，在单科上述三个步骤基础上，再加上（4）总分正态化，（5）总分标准化，（6）总分标准化后的线性转化，最终得到单科成绩和总分在 100 到 900 之间的分值。从统计与测量的角度来看，标准分相比原始分具有明显的优势。在当时的改革背景下，标准分统一了量纲，在解决不同科次考试的可比性、不同科目成绩的可加性以及提高考生志愿填报的精确性等方面，取得了实质性的进步。但其不足也比较明显，如理论模型比较复杂，计算精度要求高，需要对一些极端情况进行特殊处理，等等。在目前改革的要求下，标准分制度可能并不具备重新启用的客观条件，或者说可能带来一系列问题。

（二）引入标准分可能带来的问题

从根本上看，分数转换是一个统计分析方面的技术处理过程，对其优劣的讨论离不开对现实约束条件的关注。本轮高考改革的制度设计为分数转换设置了两个基本的约束条件，一是选科科目的多样化组合，二是坚持统一划线、总分录取的基本模式。在此，统一划线与总分录取又给不同科目之间分数的可比性、总分的合成提出了强制性的要求。

1. 在低分段和高分段出现“跳跃”，丧失一部分距离信息

由于高考成绩经常是偏态分布，因而需要对考生成绩事先进行正态转换，使得原始分服从或接近正态分布。这实际上是使用复杂的统计技术来调整成绩的分布状态。强行改变某个学科的原始分成绩分布是否有足够的理论依据，一直以来都有很大争议。例如语文学科，考生得分过高或过低的可能性一般比较小。笔者多次以省为单位来对全体考生不同科目的高考成绩进行分析，结果表明，数学、英语的标准差一般是语文的两倍。如果标准差小是语文学科的天然属性，那在标准分转换中强行将其标准差与其他学科拉平，显然有违学科的客观规律。退一步讲，即使原始分服从正态分布从而不需要进行强行正态化的话，据参与我国 80 年代标准分改革的专家估计，标准分在[- 4, + 4]之间取值的概率达 0.99993666，绝大多数考生的高考标准分 T 值在 100 到 900 之间取值（黄光扬，1997）。尽管这个概

率值很高,但一些省份的考生超过 50 万,这就容易使得正态化转换方案在低分数端和高分数端中丧失一部分在原始分视角下看来非常重要的距离信息。为验证这一情况,笔者采用当年的标准分转换程序,选取某省 A 科目的学业水平考试数据,[⑤]将原始分与转换后的标准分对应关系做成下图(见图 2)。

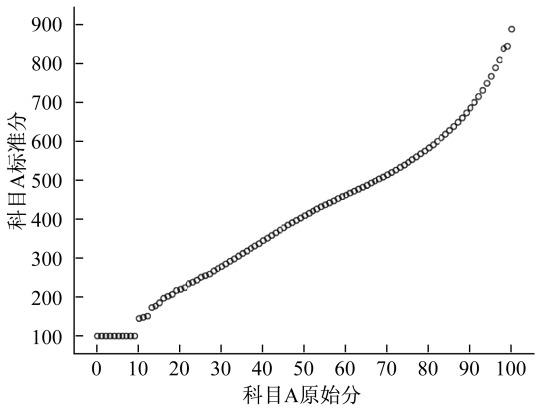

图 2 科目 A 的原始分与标准分对应关系

该科目的原始分均值 72.49,标准差为 20.04,从图 2 中可以看出,原始分转换为标准分后,标准分的区间范围从 100 到接近 900。在原始分的低分段和高分段的一些位置上,出现了间距不同甚至断点的现象,这是因为位于原始分两端的分数点上的人数较少,转换成标准分后其间距被强行拉大了。这对于处于高分段的学生来说是难以接受的,尤其是在与原始分进行比较时,高分段的原始分几分之差就会造成标准分十几分甚至几十分的差异。也就是说,为了去填标准分中的有关分数点,原始分的距离信息被扩大了。这一现象在标准分制度改革实施以来一直是个棘手的难题。

2. 选科差异削弱标准分的适用性,并将加剧学科之间的不均衡

在原始分的正态转换过程中,成绩的频次、考生群体数量都会对标准分产生一定影响。这实际上暗含了标准分使用的一个隐含条件——同一群体或人数相近的考生群体。总体上看,标准分的适用性与考生群体数量的差异成反比。如果考生群体人数过小,转换后的标准分最高分可能远远达不到设定的满分,这实际上缩短了这些科目总分的全距。同时,标准分之间的间距也会比较大,甚至不少分数点会出现空缺。这导致在总分合成时,一些考生人数少的科目会处于非常不利的地位,选择这些科目的考生则会在不同程度上吃亏。回到上述广西“3 + X”的科目改革,2003 年的科目组合有 12 个,不同科目的选考人数一下子形成了较大的差异,标准分的不足就立即暴露了出来,合理性受到质疑。因而,2004 年不得不把科目组合减少一半,以减少部分科目选考人数过少给标准分分数转换所带来的消极影响,尽管实际上仍旧是于事无补。

另外,在选科多样化的要求下,学科之间在内容、考试难度等方面的差异,考生的禀赋、兴趣以及专业性向上的差异,都会导致不同科目的考生数量出现分化。这时,标准分受群体数量影响的特性,将加剧学科之间的不均衡以及考生在选科上的博弈。在当年实施“3 + 文综或理综”科目组合改革的省份中,除了必考科目外,学生只能在文综或理综之间进行选择,这样,使用标准分造成的科目之间的不均衡也只体现在这两科之间。在实施“3 + 综合 + X”的广东省,每年的 X 科的考生人数不一。广东是第一个实施标准分改革的省份,时间最长,经验也最丰富,但这一科目改革客观上仍旧形成了“考生人数越多,越容易获得较高标准分”的局面。由此导致几年内广东报考物理的考生由十多万人迅速降至五六万人,很多考生被迫弃理从文,或选择试题难度低的科目(黄晓慧,2013)。这种状况与当前浙沪两地新高考改革试点后出现的“物理选考人数下降”“选科文科化倾向”非常相似。进一步而言,当前六(七)选三的选考方案在科目组合上比以往的改革又向前迈进了一步,浙沪两地选科科目的组合分别有

35 和 20 种，组合的增加必然导致各科目人数差异的增大。对于考生数量较多的省份，如果各科考生人数相当，标准分可能还有一定的可行性。但对于考生数量较少的省份，如今年进入新一轮高考改革试点的北京、海南来说，使用标准分可能隐藏着巨大的风险。⑥

3. 分数合成中可能引起学生位次的变动

通俗地讲，分数合成是指将各个科目成绩以一定的方式加总，以满足高校录取需要的分数处理方法。目前，由于统计、测量技术在大规模教育考试中的使用，量表分（Scaled Score）得到了广泛使用。事实上，国际同行几乎在所有重要考试中不再直接使用原始分数（形成性测试或诊断性测试除外）（杨志明，2015）。在保持总分录取的前提下⑦，各科的分数合成不得不面对一个核心问题——可加性。无疑，不同科目的原始分由于试题难度、标准差等指标的差异，直接相加不尽合理。标准分在解决单科成绩的可加性上前进了一步，但带来的问题是合成后的综合分（也就是标准化后的高考总分）会使考生的位次发生一定变化。这里的变化有两种情况，一是在原始分中达到某批次线而在标准分中没有达到。二是在原始分中达不到某批次线而在标准分中可以达到。一些推行过标准分的省份进行过这方面的分析，结果表明：原始分与转换后的标准分之间的相关达到 0.98（扈涛等，1994，第 68 页）。这个相关的值尽管比较高，但由于考生总数动辄数十万，位次受到影响的考生实际上并不少。其他分析也表明：过录取线的考生中，文史类、理工类和外语类各有高分段前三名考生的原始总分位置在标准分转换后未发生变化，其他考生有不同程度的变化（陕西省标准分改革专家组，1995，第 94 页）。这种位次变化，在传统的志愿填报模式下不那么引人瞩目。但随着平行志愿投档录取模式的推广，这一弊端引起越来越多的反对之声。平行志愿投档的基本原则是“分数优先”，在这种录取模式中，考生的位次在录取中起着举足轻重的作用，位次的任何微小变动都是极其高利害的。在高考成绩分布的一些密集位置，“一分一操场（人）”的现象正是这种情形的生动写照。目前，我国已经全面实行平行志愿录取投档模式，在这一前提条件的限制下，高考总分合成方式的变化，可能面临一定的社会风险。

三、改进的可能

（一）客观认识等级分赋分方式的必要性及其不足

到目前为止，新高考方案实施过程中出现了一些不符合育人导向的博弈行为，尤其是在选科上，不同水平的学生、不同层次的学校或多或少都卷入其中。物理科选考人数持续下降只是一个很突出的表象，深层次的一个重要原因是现行的等级赋分方式。在此，笔者认为不能简单地对个体的博弈行为加以指责。制度设计的漏洞不能让个体的理性选择来承担责任，更不让个体利益权衡下的趋利避害行为成为深化改革的阻力。从这个意义上来讲，回归方案本身的完善是理性的选择，选考科目的赋分方式也的确具有改进的空间。但在改进之前，我们需要深刻认识等级分的不足、使用的初衷以及所受的限制。

受制于多种因素，这一轮高考改革的制度设计，在招生端，仍旧坚持了总分录取的基本做法，但在考试端，科目组合呈现出前所未有的多样化。前者要求考生的分数在同一个量表上可比，以满足录取的需要。而后者由于不同科目在教学内容、考试难度等方面存在的差异，又天然存在不可比性。对这一难题可以采取的解决办法是对考试分数进行技术处理以尽可能满足可比性。实际上，等级赋分是一种介于原始分数和标准分数之间的计分方式（宋吉祥等，2017）。其思路和目标与标准分大致相同，即通过相应的技术处理，使得不同学科的成绩具有相同的分布、难度和标准差⑧（浙沪两地等级分的分数分布见图 3、图 4），以达到不同科目在分数使用上的等效⑨，最终能够得到一个相对可比的总分。

因此，从解决不同科目成绩的可加性这个单一目标来看，等级赋分无疑是一种进步，如果要彻底解决可加性，标准分无疑是更好的选择。但是，标准分在进一步提升不同科目成绩可加性上的贡献，远远不足以抵消其上文提及的种种弊病。在分数转换过程中强行正态化的做法并不比按比例划分等级的做法高明多少。因此，在新高考方案中放弃标准分，未免不是一个明智的选择。

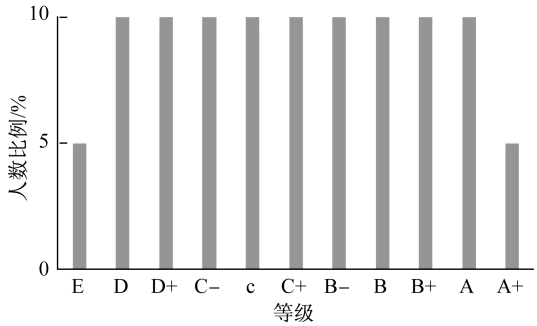

图 3 学考等级分的分数分布(上海)

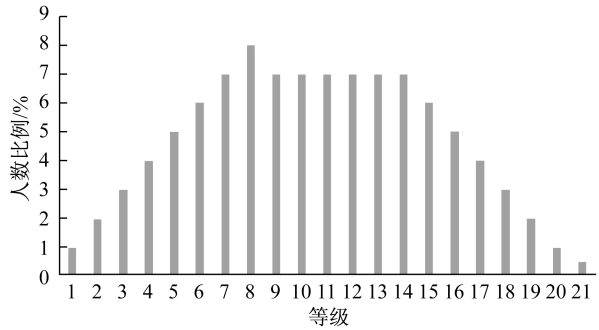

图 4 学考等级分的分数分布(浙江)

等级分是为解决不同科目成绩的相对可加而采用的一个妥协办法,但这不足以说明其自身因此就有足够的科学性。实际上,等级分的不足非常明显,其按照比例来“切割分等”的处理方式,使得考试成绩在很大程度上取决于一个与测量没有关系的外部因素——考生人数。从结果来看,等级分事实上已经造成了考试成绩既取决于“考生水平有多高”,又取决于“考生和谁一起考”的局面。这给考生带来一种不确定性,使得他们对考试的结果感到难以预期,进而营造出焦虑的氛围,并加剧他们的博弈行为。从技术细节上看,等级赋分在一些分数段上会扩大或缩小原始分之间的差距。在划分比例的边界上,很可能出现原始分相差 1 分而等级分相差 3 分的情况。也有可能出现原始分相差若干分以上,而由于分数分布形态的原因被划入同一个等级进而得到相同分数的情形。局部分数转换的偶然性与总分录取上“分分必究”的高利害性是一个难以调和的矛盾,这迫切需要回归高考的核心价值宣誓——公平的维度上来加以改进。需要指出的是,不少研究者提出对总分录取进行改革,这可能是未来的一个努力方向。笔者认为,总分录取是在对社会现实充分考量基础上,为保障高考公平而坚守的一条底线。对于这一经过实践检验并最终被社会所接受的方式,短期内恐怕不具备取消的条件。从认识论的角度来看,总分录取遵循的是实践逻辑,这是一种特定条件限制下的“在实践者与环境相互作用的历史活动中‘生成’的逻辑”(冯向东,2012),具有高度的情境关联性。尽管很多时候看上去与理论逻辑的要求不相符合,但却是诸多社会历史条件约束下的必然选择。

(二)在考试技术层面对等级分赋分方式的再审视

在教育领域,考试是对学生进行评价的一种重要手段,其本质是通过一定的方式来推测考生的心理特质。这种心理特质是内隐的、潜在的,对其进行测量与物理测量有着本质的不同。从操作上来看,试题是最小的测量单位,考试就是用试题来收集考生的作答反应并对其进行处理、赋值的过程。赋值的结果一般用具体的分数来表示,代表了考生在某个测量目标上的水平。其中,对考生的作答反应可以进行多种技术处理。原始分就是其中一种最简单的方法。从教育测量的视角来看,原始分是依据经典测量理论来对考试结果进行标示的一种方式,其计算过程不考虑试题的难度。在具体操作上,只需要依据事先确定的每个试题的分值,再根据考生作答反应与参考答案的符合程度进行简单计算,从而得到一个分值。而整卷的得分只需要将每题的原始分相加即可。在此,原始分的计分方式仅仅是在测量水平的差异,排出等级或顺序,既没有相等单位又无绝对零点,是一种典型的顺序数据,并不具备可加性。这样一个在考试技术上有瑕疵的计分方法,尽管长期以来一直被广泛采用,甚至成为一种传统,但是其不足也非常明显:一是忽视了试题难度对得分的影响。二是把顺序数据当作等距数据来使用。在总分合成过程中,不同科目原始分相加时同样不考虑科目之间的难度差异,仍然把原始分作为等距数据来进行处理,这忽视了不同科目分值的“含金量”存在的巨大差异。

总分录取这个特殊的分数使用方式决定了原始分是等距数据的基本属性(至少在功能上),这一特点已广为社会所接受,成为一项体现高考公平的共识。因此,在新高考改革方案中,必考科目的计分仍

旧使用原始分，在一定程度上体现了对这一共识的认可。但随之而来的问题是：选考科目的分数如何加总并保证不同组合之间的可比性。为了使得选考科目的成绩具有相对的可比性，人们不得不对原始分进行等级的转换，转化过程是按照百分比来对原始分进行分等，本质上是百分位数。因而，在给不同的等级赋分后得到的选考科目成绩又成为了顺序数据。这样的处理没有考虑科目难度，处理后的所有选考科目的分数分布完全一致，必然与原始分的分布形态大相径庭。

选择性是本轮新高考改革的一个亮点，总分录取所要求的可比性迫使选考科目的计分方式在科学性上做出必要的让步。实际上，选考科目的原始分经过等级赋分后成为一种简化或者不是那么平滑的标准分，这一方法回避了标准分转化中复杂的技术处理过程，而是通过对原始分进行分段划分，来使得其分布呈现出类似的正态分布（尽管在处理方法上粗糙一点），并以此来实现不同选考科目得分的相对可加性。总之，相较于原始分，等级赋分使得不同科目分数相加具备了一定合理性，但付出的代价是各科的难度、标准差被强行处理成一样，并使得分数分布扭曲。

可见，在从原始分到等级分以及合成总分的过程中，分数的“属性”发生了不少变化。现有高考科目计分方法的设计，充分认识、尊重使用原始分的传统和必要性，但也不可避免地承袭了其不可取之处。总分录取这一特殊的分数使用方式在短期内难以突破，在这一前提下为确保不同科目成绩的可比性而不得不改变数据性质的处理方式，是一种迫不得已的妥协。

（三）改进的可能思路

从实践逻辑的角度来看，等级分的技术处理是为了满足总分录取这一特殊效用而采用的一种相对简便的分数转换方式，它得到的是一种在录取上可比的等效分数（Equivalent score），而并非考试技术上讲的等值分（Equated score）。这种等效分数的得出更多地表现为分数使用上的一种约定，它可能并不遵循严格的理论逻辑，甚至看起来还有明显不足。例如，在大学以及研究生课程学习的评价中，GPA就是一种典型的等效分。表1是国内某985高校多种课程成绩评定结果之间的转换关系。从考试技术角度来看，在某个分数区间内，原始分之间的差异被抹平了，这也是一种分数扭曲，而且也不公平，其结果同样具有高利害性，因为国内外大部分高校把GPA作为评优和人才选拔的重要依据。但是，在实践中，这种成绩转换方式早已成为通行的做法，广为教育界和社会所认可。从这个角度来看，等级分的赋分方式在方向上具有一定的实践理性与民意基础。也就是说，我们可以而且也有必要保留等级分的整体框架，在局部的技术细节上来加以完善。

表1　某高校不同课程成绩评定结果的对应表

课程成绩	成绩等级	课程绩点	百分制成绩
优	A	4	100～90
良	B	3	89～80
中	C	2	79～70
及格	D	1	69～60
不及格	F	0	59～0

不管是从学理还是常识的角度来看，在现行的各种限制条件下，等级分的改进需要尽可能解决两个根本性的问题，一是考生成绩在一定程度上取决于考生群体数量的状况，二是分数转换中“忽高忽低”的现实[10]。按人数比例划分等级带来的根本性问题是等级划分标准的不稳定。因而，对于前者，统计与测量的有关技术可以解决，即通过标准设定与维护等技术手段，使得等级划分标准摆脱考生人数等外在因素的影响，以确保各个等级的标准在测量误差范围内的相对固定。这方面，欧美和我国香港地区的有关考试都已经积累了不少经验，实际上，国内其他一些社会影响力相对较小的考试也早已采用类似方法，但这些做法却很难在高考这种高社会关注度、高利害的考试中使用。对于后者，则需要尽可能保证转换过程中分数属性的相对一致，并使得转换前后的分数在两个量表上同步映射。也就是说，分数所包含的距离信息需要在转换过程中尽可能保留，可以同时缩小或放大，但绝不能出现同一等

级区间内某个局部缩小,另一个局部又放大的情况。因为分数的距离信息与考生的位次息息相关并从根本上决定录取与否。具体的改进举措是采用线性转换的方法与技术,使得各个等级区间内分数变化的趋势在转换前后保持一致,这是在分数处理层面维护考试公平的必然要求。这又反过来要求对转换前分数的区间划分不宜过多,否则将难以保证转换后分数的区分度。总之,在原始分的分数分布上设置若干关键分数点,形成若干等级,在各个等级内进行线性的分数转化,可能是一个可以尝试的思路。

最近,为避免物理选考人数的持续下降,浙江提出了"选考科目保障机制"的举措,这是出于整体的、长远的利益考虑而对这一学科给予特殊处理并使用政策工具进行调整的思路,是走出"个体理性选择导致群体非理性结果"困境的出路,短期内应该会发挥积极作用。从长远来看,选考科目的赋分方式还需要改进。在高考改革的历程中,改革的历史传承、社会的认知水平和容忍度、利益相关者的特殊诉求等,构成了诸多特定的限制条件。本文是在这些限制条件下,在保持现行赋分方式不发生根本性变化的前提下,在考试技术层面提出改进思路,或者说仅仅是一个思考框架。希望随着这一轮高考改革在更大范围内的推广或随着一些限制条件的打破,更好、更有针对性、更具体的解决方案能够涌现出来。

(本文仅代表个人观点)

参考文献

戴海崎等.(2006).*心理与教育测量*.广州:暨南大学出版社.

冯向东.(2012).教育科学的理论与实践逻辑——关于布迪厄"实践逻辑"的方法论意蕴.*高等教育研究*,(2),13 – 19.

郭小川.(2002).带着读者的疑问听瞿司长讲那高考的事.*高校招生*,(2),6 – 7.

扈涛等.(1994).*标准分制度及其应用*.北京:科普出版社.

黄光扬.(1997).高考标准分的线性转换与正态化转换.*统计教育*,(6),18 – 19.

黄晓慧.(2013 – 12 – 04).标准分,还能挺多久.*人民日报*,(12).

陕西省标准分改革专家组.(1995).*标准分及其应用*.西安:陕西省考试管理中心.

宋吉祥等.(2017).三种典型分数分布形态下等级赋分方式的比较.*考试研究*,(1),76 – 84.

杨志明.(2015).高考原始分合成:问题与改进思路.*教育测量与评价*,(10),63 – 66.

章建石.(2016).一项公平与效率兼备的高考改革为什么难以为继?——标准分制度的变迁及其折射的治理困境.*北京师范大学学报(社会科学版)*,(1),31 – 41.

注 释:

①如果完全放开,文、理类"X"将一共有 21 个组合。

②这方面包括:高校难以提前一年确定专业的选考科目要求;担心细化了科目要求影响生源选择范围,进而可能影响生源质量;有不同组合的高考总分参与录取排序,难以进行区别。

③关于标准分转换的技术细节,在笔者的另一篇论文中有介绍(参见章建石,2016),不再赘述。

④不少研究者认为,在高考中,由于考生数量庞大,可以将成绩作为正态分布来处理,这实际上是一种强假设,与事实不符合。

⑤原始分满分为 100。

⑥两地 2017 年的报考人数在 6 – 7 万之间,仅为一些报考大省(河南、广东、山东)的十分之一。

⑦总分录取在保障形式公平上具有重要的价值,短时间内很难取消。

⑧根据浙沪两地给出的等级计分方法,可以计算出选考科目的标准差,浙江为 13.8,上海为 8.75。

⑨有研究者提出这是不同科目之间的等值,这是不严谨的。

⑩这是一种形象的说法。主要是指现行等级分的赋分方式使得在不同的分数段中,有的成绩在转换后偏高,有的成绩在转换后偏低。

(责任编辑 胡 岩)

Comprehensive Quality Evaluation in the New CEE Context: Implication, Implementation and Application

LIU Zhijun ZHANG Hongxia WANG Hongxi WANG Ping WANG Hongwei
(College of Education Science, Henan University, Kaifeng Henan 475001, China)

Abstract: In the New CEE (college entrance examination) context, the application of the comprehensive quality evaluation in colleges and universities enrollment reflects not just the spirit of the reform, but also the need to promote the reform process. Based on the understanding of comprehensive quality evaluation, this article focuses on why and how the evaluation should and can be used or implemented, what risks are involved, and what guarantee conditions are needed. Then, in a systematic way, it deals with the values and implications, international experiences and practical feasibility, programming and implementation, possible risks and guarantee conditions. In conclusion, the article attempts to respond theoretically and practically to the application of comprehensive quality evaluation in colleges and universities enrollment.

Keywords: the comprehensive quality evaluation; colleges and universities enrollment; preliminary screening; re-test; risk analysis; guarantee conditions

Theory, Practice and Prospect of University Comprehensive Assessment Admission: A Case Study of Shanghai New Gaokao

TIAN Aili YAN Lingyan
(Faculty of Education, East China Normal University, Shanghai 200062, China)

Abstract: It's required that students' comprehensive quality assessment should be included as a reference in university admission, according to China's new university entrance examination reform. In theory, comprehensive quality means individual's ability to solve complex tasks in specific contexts, which is characterized by integrity, comprehensiveness and individuality. The assessment of students' comprehensive quality should focus on how they solve problems in real life. At the policy level, comprehensive quality is divided into different dimensions and high schools are required to keep a record, which will be used for reference in the admission process. In practice, on the other hand, some universities in Shanghai are exploring how to assess the comprehensive quality by means of university activities, referring to the report by high schools and examination score to improve the reform. To strength capacity-building of admission, universities should strive to improve the professionalism of assessment by providing sufficient resource in support of the admission processes.

Keywords: comprehensive quality assessment; Shanghai; self-enrolment

The Improvement of Percentile Band Score: Experience, Restriction and Possible Direction

ZHANG Jianshi
(The National Education Examinations Authority, Beijing 100084, China)

Abstract: The score comparability of different subjects in Gaokao reform has always been a big concern in recent years. Experience in some provinces shows that standard score reform is not a proper solution to this

problem because of its technical flaws in the process of score conversion. Percentile band score, despite some restrictions like using total raw score in admission, seems a compromised choice. To some extent, percentile band score, as a kind of simplified standard score, makes it possible to compare the scores of different subjects before adding them together while leaving some negative influence on teaching and learning. However, technically, percentile band score still needs improving given the existing practice of using Gaokao scores.

Keywords: Gaokao; conversion of score; percentile band score

How to Make the Results of Academic Evaluation More Valid: Research on Adjustment Model Based on Latent Variable

LIU Hui[1] ZHANG Peng[2] PAN Jingjing[2]

(1. College of Education, Zhejiang University, Hangzhou, 310028, China; 2. School of Mathematical Sciences, Zhejiang University, Hangzhou, 310021, China)

Abstract: The validity of Gaokao (Chinese college entrance examination), a selective test of academic evaluation, depends on its identification of the variability of students' transferable ability in problem-solving. However, the raw score in academic evaluation does not reflect the actual level of students´ability, which is a latent variable. In order to make the results of academic evaluation more valid, this study constructs a moderated model based on latent variable by treating a student with full score in ability as a reference. Raw score of a particular question is re-weighed according to the difficulty of the question. The moderated model based on the latent variable was applied to the data analysis of an 11-school-league examination, with a total of 9,008 high school students participating in 10 subjects tests. The results show that: a) the adjusted score is more normal than the raw score; b) the ability score is more stable than the raw score; c) the total score has a high correlation with the ability score; d) individually, there is a big difference between the raw score and the ability score.

Keywords: academic assessment; reform of Gaokao; validity; transfer; ability; latent variable

Colleges and Universities as the Subject in Enrolling Students

QIN Chunhua

(1. Graduate School of Education at Peking University, Beijing 100871, China; 2. the Institute of Examinations at Peking University, Beijing 100871, China)

Abstract: The new system of college entrance examination differs from the old one in that colleges and universities can act as the subject in enrolling students, This change from passive admission to active enrollment enables college admission criteria as a guideline to promote students' healthy growth. If admission agencies in colleges and universities cannot transform their functions and still enroll students based only on scores, the current "exam-oriented training" and "only test scores matter" will likely be further aggravated. And the pre-reform dead hand in admission may lead to more problems with grade assignment, the conversion and comparison of multiple calculation methods, equivalence and so on. On the other hand, if the admission agencies transform their duties by truly implementing the spirit of *two basics and one reference*, it is likely to achieve the goal of *strengthening moral education* and that problems like profit-seeking in the pilot reform can also be easily solved.