标准参照测验分数体系的探讨研究

甘良梅　余嘉元
(南京师范大学 心理系,南京 210097)

摘　要:随着考试事业的发展,标准参照测验(Criterion Referenced Test, CRT)也越来越多地受到人们的关注,但是它却陷入了用常模参照测验的方法来解释和报告分数的误区。该文从国内外重大标准参照测验 CET－4 & CET－6、HSK、GRE、CLEP 等分数体系入手,通过对其分数体系的共同点分析,探讨出适合于标准参照测验的分数体系,最后指出目前一些测验的分数体系仍然存在的问题。

关键词:标准参照测验;分数体系;常模参照测验;测验等值;分界线
中图分类号:B841.7　　文献标识码:A　　文章编号:1003－5184(2006)03－0079－04

1　提出问题

1.1　标准参照测验的定义

1962 年,匹斯堡大学的 Glaser 首先提出了标准参照测验(Criterion Referenced Test,CRT),并把测验分成了标准参照测验和常模参照测验(Norm Referenced Test, NRT)。[1] 但是 Glaser 对标准参照测验给出了一个宽泛的定义,认为标准参照测验"是一种精心编制的测验,它根据特定的表现标准产生可直接解释的测量,表现标准一般通过定义一个类或者一个域来说明",造成了目前都没有一个大家公认的标准参照测验定义。后来许多学者把标准参照的同义词如内容参照、领域参照等解释为标准参照。如安妮·安娜斯塔西等就将标准参照测验称为领域参照测验,认为此种测验的解释参照系是规定的内容领域而不是规定的人员总体。[2] 国内曾桂兴将标准参照测验称为"目标性考试",即认为"依据考试的既定目标的标准来解释分数的考试,用以描述考试所达到既定目标的程度"。[3]

标准参照测验因其侧重于测验的内容意义以及测验的掌握水平等方面,而得到广泛的运用:诊断是否达到了必要的技能、调查大规模的教育考试,以及一些专业技能的测试,如评定是否取得驾驶执照或者达到军事技能的水平等。[2]

1.2　目前标准参照测验的误区

但是在标准参照测验的统计分析上,却长期误用经典测验理论的方法,如在测验的信度估计、项目的难度和区分度、分数的报告和解释及分界线的划定等方面。而目前国内许多测验学者已经致力于前面两个方面的研究:在标准参照测验的信度估计上面,主要采用了概化理论的方法来估计其信度;[4] 在项目的分析上,运用项目反应理论的信息函数方法对试卷和项目进行分析。[5]

可是在标准参照测验的分数报告、解释和分界线的划定即分数体系方面,目前研究者关注得还不够。在分数的报告和解释上面,许多考试仍沿用了经典测验理论中的常模参照方法来对其分数进行报告和解释,即在分数报告的过程中常将分数通过 Z 分数转化得到一定的量表分数来报告。而常模参照主要依赖于其所在的样本,当样本成绩呈非正态分布将导致报告分数出现偏低或偏高的情况。另外在分界线的划定上面,即判断考生是否掌握时,依据常模参照很难确定具体区分考生掌握与否的分界线,而测验标准的划界分数(cut－off score)或者说合格线的准确把握问题,是标准参照测验编制开发中的核心与关键。[5]

因而,在标准参照测验日益被人们运用和关注时,也更加需要一个适合于标准参照测验的分数体系。研究将从国内外的重大考试入手,通过探讨其分数体系的共同点以找到适合于标准参照测验的分数体系,为以后的一些标准参照测验的分数体系提供参考。

2　国内外标准参照测验的分数体系

2.1 大学英语四、六级考试

大学英语四、六级作为一种尺度相关——常模参照测验，其客观题部分采用百分制为形式的正态分制，而主观题部分采用等级分制。在客观题部分，首先进行加权处理，然后通过建立在 IRT 基础上的等值将新测验的分数等值到标准试卷中，最后根据常模得到一个平均分为 67.8，标准差为 12.08 正态分布，而为适应我国及格线为 60 分的百分制，其对经过加权处理的原始分数做线形转换，最终得到报告分数。主观题部分主要是把语言能力分为若干等级，每个等级用明确的描述语加以规定。[6]

在其测验等值设计中，自 1987 年第一次考试以来，大学英语四、六级一直采用了 IRT 中的 Rasch 模型对同一被试组进行等值。[14] 但是由于此模型只考虑题目难度而不考虑区分度，2003 年底大学英语四、六级考试开始用双参数逻辑斯蒂模型，[7] 且用共同题目的等值设计取代原来的共同被试等值设计。其标准试卷的选择，如四级，是根据英语四级的规定“要求重点大学学生基本上（即 85% 左右）达到《大学英语教学大纲》中规定的课程内容’进行编制的，且征求全国 50 位专家的意见。标准样本是选择了六所全国一流重点大学（北大、清华、上海交大、复旦、西安交大）学生，且每所学校考生都超过 1000 人，作为实验对象。然后对他们进行施测，建立常模，最终得到标准样本。[8]

2.2 汉语水平考试（HSK）

汉语水平考试是为测试母语非汉语者（包括外国人、华侨和中国国内少数民族人员）的汉语水平而设立的国家级标准化考试。其报告分数是一种经过转换的标准分数，先将原始分数通过测验等值，得到了转化后的分数，然后再通过量表分得到了分数的等级。对于每一等级的 HSK 除了有固定的分数段，还有对其能力的具体描述。[9]

HSK 等值采用了 Tucker 真分数、Levine 真分数观察分数线形等值及等百分位模型等对测验进行等值，且在 1998 建立了基于 IRT 的题库的测验等值。2004 年 HSK 开始引入共同组设计的线形等值，对多份试卷进行了不含共同题的共同组线形等值。同时也引进了分半组卷的共同组等值方法。将新试卷的分数等值到标准试卷中。HSK 采用了 1989 年 HSK（初中等）正式开考题的试卷作为标准试卷，而由北京语言学院的一、二年级的学生组成标准样本，他们基本代表学习汉语的人的一般水平和一般的进步水平。[10]

其等级分数的划定一方面根据《汉语水平等级标准和等级大纲》所规定的对每一个等级考生的具有能力的描述，另一方面根据其等级对应的特定年级、学期和学习阶段，共同制定等级分数，在制定分数初始阶段，每一级别加上一个半级，后来到 1997 年 HSK（基础）经过鉴定，确立相应的等级分数也就最终确定。[9]

2.3 GRE（Graduate Record Examination，美国研究生入学考试）

GRE 是美国、加拿大大学各类研究生院（除管理类学院，法学院）要求申请者所必须具备的一个考试成绩，包括一般能力考试（general test）和专项考试（subject test）。一般能力考试先将各部分正确答案的数目记录下来，得出原始分数（Raw score），即答对一题给一分并且相加，然后根据 GRE General 分数转换表将原始分数转换成量表分（Scaled Score），再将各部门的比率分相加得出总分。专项考试虽无统一的分数换算规则和标准，但其分数也须转换到量表分（Scaled Score）中而被报告。

GRE 等值方法采用了三参数逻辑斯蒂模型（Logistic Model）的非随机组锚测验（外部测验）的方法，即锚题是在正式测题之外，包括需要等值的语文、数学和分析题目。且在等值过程中，为降低等值误差采用六个版本的测验，实施到七个考试组中，且在第一个和最后一个使用的测验版本一样，通过循环将测验进行等值。[11]

2.4 美国大学水平考试（College Level Examination Program，CLEP）[12]

美国大学水平考试是由美国大学委员会主办，在全美影响范围最大的一项以考试获得学分的国家考试项目。CLEP 先将原始分数（Row Score）首先转化为公式分数（从原始分数扣除受随机猜想影响而获得的分数），然后再把此分数转化为量表分数（Scaled Score），最后报告出考生成绩。对于主观题部分，如作文，其分数都是与客观题部分相结合后一起转换到量表上的而被报告出来的。对于每一个课程都有各自的量表，而通过量表分数，考生可以与适当的参考样本进行比较来说明考生的水平。

CLEP 的每门课程的成绩量表是从通过选取参考样本进行试验测试(Do Pilot)而获得。即 选取不同高校刚学完同一门课程(一般是 CLEP 科目考试的考点学校)的学生组成参考样本(the reference groups),再由他们的考试成绩进行统计分析,制定量表。而作为考生进行比较的标准样本。CLEP 考试标准的确立通过安格夫(ANGOFF)法或‘NO—YES”法,即专家评定法。CLEP 考试的及格分数线是根据其目的,即完成某课程学习的平均水平的考生能够答对试题的 50% 而确定的。其具体的分界线的划分是利用其标准样本的考试结果确定的。

3　标准参照测验的分数体系的建立

通过综述国内外的一些标准参照测验的分数体系,可以看出 不同的考试虽然其目的不一样,分数的报告形式、解释也有差异,但是他们分数的转化、等值及报告等一系列过程,均有一些共同的地方。于是可以从中找到适合于标准参照测验的分数转化报告规律,将其运用到一般的标准参照测验的分数体系设计中,从而形成适用于标准参照测验的分数体系。目前国内外标准参照测验分数体系共同点如下。

3.1　以常模为参照样本解释和报告测验分数

国内外重大的标准参照性考试均采用了一个固定的标准,即一个固定常模来对标准参照测验进行解释。对于常模参照测验,由于它存在着许多不足和缺点,如强调好的分数、反映学生的实际成绩,但未指明学生的真正的成就等。因而人们在编制一些考试测验时常常不使用常模参照,特别在 Glaser 提出了标准参照测验并对标准参照测验和常模参照测验进行区分以后,人们在使用标准参照测验时,常常尽量避免使用常模参照的方法来分析标准参照的测验,以在方法上将两者区分开。

但是,常模参照本身又有很多的优点,如考生的成绩具有可比性,即考生可以与所在的群体中其他成员进行比较。因此在许多标准化测验中,如高考等,还是会使用常模参照的方法。但是在标准参照测验中,由于此测验是使考生与固定的标准来进行比较,而考生对课程目标达到何种程度,却没有绝对的标准,一名学生的成绩高低也只有在与他人的比较中才有合乎实际的认识。因而常模参照的方法也一直被广泛地运用到标准参照测验中,另外西方一些测验学家也认为没有常模的测验不能认为是标准化的测验。因此在标准参照测验的分数的体系建立中,首先要建立一个常模样本作为标准。这样才能使考生能与具体实际的标准进行比较。但是另一方面也要注意,在标准参照测验中,虽然重视考生分数在样本中的位置,可是标准参照测验毕竟与常模参照测验不同,它是从测验的内容出发,因此在标准参照测验中使用常模参照的一个显著特点就是要求测验中的常模样本一定要具有代表性,让标准参照测验中的标准具体化,才能使得其测验成为真正的以标准为参照的测验。目前许多标准参照测验已经采用了这一做法,如在四、六级考试中,就选择了全国六所重点大学的学生作为标准样本,另外根据了《大学英语教学大纲》的规定来确定其标准 ;CLEP 采用将不同高校(一般是 CLEP 科目考试的考点学校)刚学完同一门课程的学生作为参照样本,根据他们的成绩作为其标准。

3.2　采用测验等值法转换和报告分数

许多标准参照测验报告的分数采用测验等值法,将原始分数通过标准试卷转换为量表分数,然后再报告给考生。测验等值法,即将不同测验形式的分数转换到同一分数量表上,以便不同测验形式的测验结果可以比较[13]。

由于测验等值能保证不同形式和题目的考试能进行比较从而保证了测验之间的相对稳定性。因此在标准参照测验中测验等值被广泛使用。比如大学英语四、六级考试早在其开考之时就一直进行等值处理,且其等值方案也随着时间改变。由开始的只考虑难度差异所建立的 Rasch 等值发展到如今的双参数逻辑斯蒂模型。另外,HSK 也一直运用测验等值对原始分数进行处理,并在测验等值的方法上做了大量的实证研究,结果发现在整个基于经典测验理论(Classic test theory, CTT)和项目反应理论(Item Response Theory, IRT)的测验等值法中 Tucker 法的等值误差最小,其等值效果最佳。[14] 而在对各种基于 IRT 的等值方法的比较中,研究发现同时估计单参数的等值方法其等值误差较小。在国外大型的标准化考试中,如 GRE,自开考之初起就一直使用测验等值将每一次的测验分数等值到标准试卷中,以保证测验报告分数的稳定性。其测验的等值方案也一直在进行探索,目前仍沿用基于 Tucker 和等百分位等值法以及三参数逻辑斯蒂模型等等值方法来对

测验进行等值处理。

3.3 将常模参照与标准参照结合划定分界线

在标准参照测验中,其分数线的划定可以说是其关键和核心。目前许多的考试中均采用最低分数线来划定考生的水平。如大学英语四、六级考试中规定考生的成绩达到 60 为合格;在 HSK 考试中,无论是基础、中等还是高级,其每一个等级证书的获得都有一个设定的最低分数线。

但是在实际的操作中,如何准确地划定其最低分数线成为许多测验专家所关注的问题。由于在划分过程中,需要直接通过分数将考生划分为两类人群,即通过和不通过。但是人的能力它本身是连续性的,并不存在完全的掌握或者不掌握,只是程度的高低。因此到底在哪种能力水平以上被认为是掌握或通过就是所划定的分界线。在分数划定方法的探索上,许多大型考试采用了一些切实可行的方法。如在一些考试中,采用专家判定法中 Angoff 方法,即由专家直接判断处于临界水平的被试在某测验的每一题目正确作答的可能性,然后通过加权得到测验分数分界线。[15] 在 CLEP 考试中,让专家们根据测验本身的要求即:完成某课程学习的平均水平的考生答对试题的 50%,来确定其课程的及格线。另外一些分界线是根据一定的通过率来确定的,即根据标准样本的成绩来确定在某一通过率的临界分数。如 HSK 考试,其等级最低分数线均根据标准样本的成绩来确定的。但是更多的测验如大学英语四、六级、GRE 等,他们采用将常模参照和标准参照相结合的方法,即一方面根据测验本身的要求考虑一定的通过率,如英语四级规定"重点大学学生 85% 达到其大纲要求的内容";另一方面也根据测验所选择的标准样本的分数分布,共同决定其测验的分界线。

总之,从上述分数体系的共同点中探讨到适合于标准参照测验的分数体系,具体表现如下:首先,在分数的报告和解释上,一方面要选择一定的常模样本作为参照的标准,另一方面要保证其常模样本具有一定的代表性和标准。其次,在分数的转换过程中,采用测验等值法,而等值方法中基于 CTT 基础上的 Tucker 法、百份位数等值法以及基于 IRT 的三参数逻辑斯蒂等值法的等值误差较小。最后,在分界线的确定上,将以标准为参照的方法与以常模为参照的方法结合起来共同决定更为适宜,即一方面考虑根据测验本身的要求所确定的通过率等一些标准性的指标,另一方面也要根据所选择的标准样本的分数分布。

4 目前一些标准参照测验的分数体系存在的问题

尽管上述一些国内外重大考试采用的分数体系比较完善,如:标准样本的选择上更具代表性、等值处理的方法及其分数线的划定上也更具科学和实用性等。但是目前正在使用的一些标准参照测验的分数体系中仍存在大量的问题:

首先,分数的报告和解释过分依赖于标准样本的分布。在使用常模样本作为参照样本时就必然导致其分数过于依赖于标准样本的选择和分布。

其次,在测验等值中,采用不同的等值方法和等值设计将造成了不同的等值误差,且差异较大。但是样本容量限制等现实情况使得许多测验不得不采用一些误差较大的等值方案。

再次,标准样本和标准试卷的选择上的主观性。样本选择由于考虑到一定的施测情景,因此样本的代表性就受到了一定的限制,另外由各专家评定的标准试卷同样无法避免其主观性。

最后,分界线的划定上仍存在一定的主观性。尽管采用了专家评定方法和样本分布共同来确定分界线,但是具体的分数线的确定因人的能力是连续变量而仍具有一定的主观性。

综上所述,在目前所接触到的一些标准参照测验中,仍然存在大量有待解决的问题,它需要人们在以后的工作中进行进一步的探索,希望以后有更多更好的办法来解决这些问题,而为日后一些标准参照测验分数体系的设计提供切合实际的参考。

参考文献

1 张凯. 标准参照测验理论研究. 北京:北京语言文化大学出版社,2002.29.

2 安妮·安娜斯塔西. 心理测验. 杭州:浙江教育出版社,2001.101.

3 曾桂兴. 标准化考试常识. 成都:四川教育出版社,1987.67.

4 杨志明. 标准参照测验及其等级线信度的概化理论分析. 心理学探新,2003,(3):52-56.

5 漆书青,周骏,张青华,等. 用信息函数法对标准参照测验作质量分析. 心理与行为研究,2003,(1):34-39.

6 杨惠中. 语言能力的分级测试—大学英语四、六级考试设计中的量化分析. 考试研究,2002,(1):55-70.

7 朱正才，杨惠中，杨浩然. Rasch 模型在 CET 考试分数等值中的应用. 现代外语，2003，(1)：70 – 74.
8 朱正才. 大学英语四、六级考试分数等值研究—一个基于铆题和两参数 IRT 模型的解决方案. 心理学报，2005，(2)：280 – 284.
9 张凯. HSK 等级分数问题. 世界汉语教学，2004，(1)：71 – 80.
10 谢小庆. 关于汉语水平考试的分数体系的几点说明. 学汉语，1995，(6)：30 – 32.
11 Neal M. Kingston，Paul W. Holland. Alternative Methods of Equating the GRE General Test. GRE Board Professional Report，1986，(5)：81 – 16p.
12 柳博. 美国大学水平考试. 考试研究，2002，(1)：125 – 139.
13 漆书青，戴海崎，丁树良. 现代教育与心理测验学原理. 北京：高等教育出版社，2002.201.
14 谢小庆. 对 15 种测验等值方法的比较研究. 心理学报，2000，(2)：217 – 223.
15 戴海崎，张峰，陈雪枫. 心理教育测量. 广州：暨南大学出版社，1999.238.

The Study of Criterion Referenced Test's Score System

Gan Liangmei　Yu Jiayuan
(Psychology Department，Nanjing Normal University，Nanjing 210097)

Abstract：With the development of the test，more and more people pay attention to Criterion Referenced Testing (CRT). But based on norm referenced，there are some problems in the explanation and report of the scores. Through analyzing the score system of some tests：CET – 4&CET – 6、HSK、GRE、CLEP and following the common of these score systems，this study discuss the score system of criterion – referenced test. At last，authors indicate some problems in some test's score systems at present.

Key words：criterion referenced test；score system；norm referenced；test equating；cut – score

(上接第 51 页)

A Psychological Survey of Detection and Avoidance of Collision Events

Liu Ruiguang
(School of Education，Jiangxi Normal University，Nanchang 330027)

Abstract：Concepts of collision and braking action in driving are defined by optical information. Psychological factors and cognitive processes，which influence observers' detection of collision events，are discussed. Various psychological methods of regulation of braking action based on envirenment information，social information and individual information of collision avoidance are further studied in this paper，so that a safe motion can be obtained by drivers.

Key words：collision events；braking action；detection；avoidance