心 理 学 报 2005,37(6):832~838
Acta Psychologica Sinica

等级反应模型下项目特征曲线等值法在大型考试中的应用

周 骏[1] 欧东明[2] 徐淑媛[1] 戴海琦[1] 漆书青[1]

([1]江西师范大学教育学院,南昌 330027) ([2]人事部考试中心,北京 100080)

摘 要 在中国最大的资格考试之一的经济专业资格考试中,为保证不同年度间考试的可比性、进行题库建设和为计算机自适应考试做准备,应用项目反应理论中等级反应模型下的项目特征曲线等值法,采用锚测验等值设计,实现了4个年度考试资料的项目参数和能力参数的等值,并成功地组建了经济专业题库。在此基础上,利用等值技术对不同年份试卷的划界分数进行了比较,为经济考试的合格标准制定、确保考试的公平性提供了实证依据。

关键词 项目反应理论,等级反应模型,参数等值。

分类号 B849

1 引言

研究测验等值,使我们有可能应用测验等值的理论和方法进一步提高考试、测验的命题质量,使不同年度间试题试卷的难度具有可比性;同时,也可以对不同地区、时间的同一学科的考试分数做出可靠的比较和分析,从而确保考试、测验的公平性和考试评价的有效性。

在我国,测验等值的研究是测验研究中最薄弱的一个环节,许多重要的考试都尚未实现等值。例如,参加人数众多、影响范围很广的高等学校招生考试,自学考试,公务员考试,以及大部分职业资格证书考试等,均尚未实现测验等值。由于没有实现等值化,不同年度举行的考试的成绩不具备可比性,这样,测验的评价标准或证书授予标准就有可能受到试卷难度起伏的影响。因而,测验等值不仅仅是教育测量学研究和应用中的一个非常重要的问题,也日益成为各种考试机构关注的问题。目前将等值技术应用到实际考试的有CET(大学英语四、六级考试)[4]以及HSK(汉语水平考试)[5]。

人事部主持的经济专业考试是我国的一项重要的资格考试,开考12年以来,每年考生人数约80万。为保证不同年度间试卷的可比性、进行题库建设和为计算机自适应考试做准备,以项目反应理论为基础,运用Samejima等级模型下的项目特征曲线等值方法,实现了2000年至2003年4个年度考试资料的项目参数和能力参数的等值,并成功地组建了经济专业题库。

2 研究的方法与过程

2.1 项目特征曲线等值原理

测验等值设计有两种:一是共同被试设计("锚人");一是共同用项目设计("锚题"或称锚测验)。在项目特征曲线等值实施过程中,均使用锚测验设计。

应予等值的测验分别向不同的被试组施测,这些测验都包含了一批共同的项目或一个额外的"锚测验"。这批共用项目或"锚测验"由于向所有被试组都施测了,那么从这些被试组上获得的资料就可借助这批共同项目或"锚测验"的作用而彼此联系起来。

我们取得的实测资料,包含两部分,一个是头一年度的测验x在被试组Nx上的反应资料;另一个是次一年度的测验y在被试组Ny上的反应资料。然后,我们运用参数估计程序,对所获得的这两部分资料分别同时估出有两套量纲系统的能力与项目参数值。由于经济专业考试试卷全部由选择题构成,其中单选题部分为全或无记分方式,满分值1分,而多选题部分视选对的应选项的多寡给予不同的分数,实际评出的分数有0(全错),0.5,1.0,1.5,2(全对),即为等级记分。根据这种情况,若测验数据能支持单维性假设,我们主张采用双参数Samejima等

收稿日期:2004-11-25

通讯作者:周骏,E-mail:jxnuzj@126.com,电话:0791-8507064

级反应模型来分析处理得分资料并进行测验等值工作。

这样,若记测验 x 系统铆题的运算特征函数——即 x 系统中被试 t 在铆题 i 上恰得 j 等级分的概率(V 是共用铆题或称“铆测验”,以下含义相同):

$$P_{xvtij}=P_{xv}(\theta_{xt},a_{xvi},b_{xvij}) \tag{1}$$

同样可记测验 y 系统中铆题的运算特征函数——即 y 系统中被试 s 在铆题 i 恰得 j 等级分的概率为:

$$P_{yvsij}=P_{yv}(\theta_{ys},a_{yvi},b_{yvij}) \tag{2}$$

则被试 t 在 x 系统铆题测验第 i 题的真分数为:

$$\xi_{yvsij}=\sum_{j=1}^{k_i} j\cdot p_{yv}(\theta_{ys},a_{yvi},b_{yvij}) \tag{3}$$

进一步,被试 t 在 x 系统铆题测验上的真分数则为:

$$\xi_{xvt}=\sum_{i=1}^{n_v}\sum_{j=1}^{k_i} j\cdot p_{xvtij}(\theta_{xt},a_{xvi},b_{xvij}) \tag{4}$$

同理,可得被试 s 在 y 系统铆题测验上的真分数为:

$$\xi_{vys}=\sum_{i=1}^{n_v}\sum_{j=1}^{k_i} j\cdot p_{vysij}(\theta_{ys},a_{vys},b_{vyij}) \tag{5}$$

注意到虽然共用铆题即“铆测验”是同一批项目,但却会在两个量纲系统上有两套不同数字表现形式的参数值,所以,$a_{xvi}\neq a_{yvi}$,$b_{xvij}\neq b_{yvij}$。

按照项目反应理论,同一项目在不同量纲系统上虽然参数值的数字表现形式不同,但实质却一样。同一项目的两套参数值间必然存在以下线性关系:

$$a_{xvi}=\frac{a_{yvi}}{a} \tag{6}$$

$$b_{xvij}=\alpha\cdot b_{yvij}+\beta; \tag{7}$$

同时,按照项目反应理论,难度参数跟能力参数标刻在同一单位系统上,故 b 值上的转换关系对 θ_{xt} 与 θ_{yt} 也成立,即也有:

$$\theta_{yt}=\alpha\cdot\theta_{xt}+\beta \tag{8}$$

其中 α 是斜率;β 是截距;θ_{xt} 是被试 t 在测验 x 上的能力;θ_{yt} 是被试 t 转换到测验 y 上的能力。正因为有这样的线性关系,我们就可以把测验 x 上估出的被试能力参数值,按上式转换到测验 y 的量纲系统上去;也就可以用这些值来计算被试 t 在共用铆题上的真分数,再使用铆测验题在测验 y 上估出的项目参数值求出被试 t 在测验 y 系统上铆测验的真分数。这样,由于 ξ_{xvt} 和 ξ_{yvt} 是同一被试在同一铆测验上两种不同量纲系统上的真分数值,按项目反应理论原理两者必定相等,即有:

$$\begin{aligned}\xi_{xvt}&=\sum_{i=1}^{n_v}\sum_{j=1}^{k_i} j\cdot p_{xvtij}(\theta_{xt},a_{xvi},b\,xvij)\\&=\sum_{i=1}^{n_v}\sum_{j=1}^{k_i} j\cdot p_{yvtij}\cdot(\alpha\cdot\theta_{xt}+\beta,a_{yvi},b_{yvij})\\&=\xi_{yvt}\end{aligned} \tag{9}$$

由于式(8)只在变量为参数状态下才成立,而实际上我们只能获得项目与被试的参数估计值,故可令:

$$F=\sum_{t=1}^{N_X}(\hat{\xi}_{xvt}-\hat{\xi}_{yvt})^2 \tag{10}$$

F 为含有 α、β 及 θ_{xt},b_{xvij} 与 b_{yvij} 等参数的函数,α,β 为未知参数。这样,就可在 F 最小的条件下来估计未知参数 α 与 β,即令:

$$\frac{\partial F}{\alpha}=\frac{\partial F}{\beta}=0 \tag{11}$$

所估 $\hat{\alpha}$ 与 $\hat{\beta}$ 即为等值常数。有了等值常数,我们就可以根据式(6)、(7)和(8),进行项目反应理论中的项目参数等值和真分数等值及观察分数等值。

上述这种等值方法就是项目特征曲线等值法,是由黑巴拉、斯托金和洛德所提出。[1,2]

大量的研究已经说明,在模型—资料拟合检验支持所采用的反应模型是合适的情况下,项目反应理论等值要优于经典测验理论等值;在采用项目反应理论等值方法时,项目特征曲线等值法又要优于一般的平均数标准差法和稳健的平均数标准差法等值,因为项目特征曲线等值既承认作为基础数据的项目与被试参数是估计值,又能全面利用难度与区分度参数所提供的信息[1,2]。所以,我们在自己的测验等值工作中就坚持采用这一方法。为此,开发编制成了等级反应模型项目特征曲线等值专用计算机程序,并将它作为专门模块纳入江西师大为人事部考试中心开发编制的“经济专业题库”分析计算系统中。

2.2　Samejima 等级模型介绍

对双参数 Samejima 等级反应模型(Graded Response Model)做简单的介绍。

我们先把双参数 Logistic 函数模型写成下面的形式:

$$P_{i1}^{*}(\theta)=\{1+Exp[-1.7a(\theta-b_{i1})]\}^{-1} \tag{12}$$

这个模型适用于等级数 $k_i=1$ 的评分资料,也即适用于(1,0)记分的项目。P_{i1}^{*} 的意思是:能力为 θ 的被试在项目 i 上评为 1 等的概率为 $P_{i1}(\theta)$。由于没有第 2 个等级,因此我们也可以理解为,能力为 θ 的被试在项目 i 上评为 1 等及 1 等以上的概率为

$P_{i1}^*(\theta)$。再把上面的函数模型拓展理解，我们可以写出能力为 θ 的被试在项目 i 上评为 0 等及 0 等以上的概率为：

$$P_{i0}^*(\theta) = \{1 + Exp[-1.7a(\theta - b_{i0})]\}^{-1}$$

同样可以写出能力为 θ 的被试在项目 i 上评为 2 等及 2 等以上的概率：

$$P_{i2}^*(\theta) = \{1 + Exp[-1.7a(\theta - b_{i2})]\}^{-1}$$

对于记分方式为(1,0)的项目来说，被试在这样的项目上作答，被评为 0 等的难度为 $-\infty$，故评为 0 等及 0 等以上的概率 $P_{i0}^*(\theta) = 1$；而被试被评为 2 等的难度为 $+\infty$，那么评为 2 等及 2 等以上的概率 $P_{i2}^*(\theta) = 1$。由此，我们可以认为一个单等级的项目可以由下面三条概率曲线描写：

$$P_{i0}^*(\theta) = 1$$

$$P_{i1}^*(\theta) = \{1 + Exp[-1.7a(\theta - b_{i1})]\}^{-1}$$

$$P_{i2}^*(\theta) = 0$$

每一条概率曲线表示被试在项目 i 上得 l 等及 l 等以上的概率。利用这三条曲线，我们可以求出能力为 θ 的被试在项目 i 上恰得 0 等的概率 $P_{i0}^*(\theta)$ 和恰得 1 等的概率 $P_{i1}^*(\theta)$ 如下：

$$P_{i0}(\theta) = P_{i0}^*(\theta) - P_{i1}^*(\theta) = 1 - P_{i1}^*(\theta)$$

$$P_{i1}(\theta) = P_{i1}^*(\theta) - P_{i2}^*(\theta) = P_{i1}^*(\theta) - 0 = P_{i1}^*(\theta)$$

如果项目 i 共有 k_i 个等级，对于任一个等级 l ($l = 0,1,2,3,\cdots\cdots k_i$)，我们把全体被试作一个(1,0)式划分：凡评得等级数在 l 等或 l 等以上的被试记为 1，评得等级数在 l 等以下的被试记为 0。根据这种在 l 等上的划分，可以按 Logisitc 函数形式写出被试在项目 i 上评得 l 等及 l 等以上的概率为：

$$P_{il}^*(\theta) = \{1 + Exp[-1.7a(\theta - b_{il})]\}^{-1}$$

加上得 0 等及 0 等以上的概率：$P_{i0}^* = 1$

和得 $k_i + 1$ 及 $k_i + 1$ 以上的概率：$P_{i(k_i+1)}{}^*(\theta) = 0$

总共有 $k_i + 2$ 条得某等及某等以上的概率函数曲线。

$$P_{i0}^*(\theta) = 1$$

$$P_{i1}^*(\theta) = \{1 + Exp[-1.7a(\theta - b_{i1})]\}^{-1}$$

$$P_{i2}^*(\theta) = \{1 + Exp[-1.7a(\theta - b_{i2})]\}^{-1}$$

$$\cdots\cdots$$

$$P_{ik_i}^*(\theta) = \{1 + Exp[-1.7a(\theta - b_{ik_i})]\}^{-1}$$

$$P_{i(k_i+1)}{}^*(\theta) = 0$$

利用这组类型特征曲线，我们可以求得能力为 θ 的被试在项目 i 上恰得 l 等的概率函数 $P_{il}(\theta)$ ($l = 0,1,2,3,\cdots\cdots k_i$)。

$$P_{i0}(\theta) = 1 - P_{i1}^*(\theta)$$

$$P_{i1}(\theta) = P_{i1}^*(\theta) - P_{i2}^*(\theta)$$

$$P_{i2}(\theta) = P_{i2}^*(\theta) - P_{i3}^*(\theta)$$

$$\cdots\cdots$$

$$P_{ik_i}(\theta) = P_{ik_i}{}^*(\theta) - P_{ik_i+1}{}^*(\theta) = P_{ik_i}{}^*(\theta)$$

这就是 Samejima 导出的等级评分资料的数学模型(Graded Response Model)。要注意的是等级反应模型中每一条概率曲线都是两条类型反应特征曲线之差，新的概率曲线组被称为运算特征曲线[1~3]。

2.3 研究对象

等值试卷为人事部经济专业资格考试 2000 到 2003 这 4 年的初级与中级试卷(各 4 份)。2000 年初级与中级试卷各有 90 题，其他年份的初级与中级试卷各有 105 题，每份试卷含有 15 道铆题，即同一级别的相邻两年的试卷上分别有 15 道题目是相同的。被试得分矩阵为从当年(2000 年到 2003 年)全国随机抽样获得的 12000 考生的实测资料。

2.4 分析工具

江西师范大学教育与心理统计测量研究开发中心研制的“经济专业题库系统”(PEB)的等值专用模块；PARSCALE 4.0 以及 SPSS 10.0 软件。

3 研究结果与分析

3.1 单维性检验

前已说明，使用项目反应理论进行等值的前提条件是测验反应数据必须符合单维性假设，一般认为当采用因素分析方法抽取的因素，汉普尔顿具体提出如果第一因素特征值大于第二因素特征值近于 3 倍或以上时，可认为数据是单维的[1,2]。因此，我们在做等值工作之前，对这些年份的数据进行了因素分析，结果见表 1。

表 1 四个年份测验单维性检验结果(表内数据为特征值)

年份	初级		中级	
	第一因素	第二因素	第一因素	第二因素
2000	7.085	2.403	7.113	3.164
2001	5.19	2.7	7.297	2.916
2002	11.893	2.9	12.2	3.23
2003	12.553	2.551	13.603	2.810

表 1 结果表明，2000 年与 2001 年测验的单维性不见得很有保证，但第一因素的特征值也都是第二因素的 2 倍以上。2002 年及以后的测验单维性则都是能得到切实保证的。

3.2　模型—资料拟合检验

模型－数据的拟合性检验是针对所选模型与实测的考试数据进行的,目的是检验所选模型与其处理的资料之间是否匹配,以确定模型应用的有效性。项目反应理论中,等级反应模型的拟合检验与单等级(全或无记分方式)模型的拟合检验不同,等级反应模型的拟合检验采用的检验指标称为似然比卡方统计量(likelihood-ratio chi-square statistic)[11],近似服从自用度为$(H-1)(m-1)$的卡方分布。公式如下:

$$G_j^2 = 2\sum_{h=1}^{H_j}\sum_{k=0}^{m_j} r_{hjk} ln \frac{r_{hjk}}{N_{hj}P_{jk}(\bar{\theta}_h)}$$

其中,G^2 为似然比卡方统计量;j 为第 j 个项目;m_j为项目 j 的满分;k 为第 k 个等级分;H_j为在第 j 题上将估计出的所有被试能力参数 θ 划分为 H 个区间;r_{hjk}为项目 j 在第 h 个区间,第 k 等级分上答对的人数;N_{hj}为在项目 j 的第 h 个能力区间上所有被试个数;$\bar{\theta}_h$ 为第 h 个能力区间上被试能力的平均数;$P_{jk}(\bar{\theta}_h)$为在第 h 能力区间,第 k 个等级分上平均能力的概率值,使用式(12)计算。

我们使用此方法,随机抽取了某年中级考试的2000 人进行分析,该份试卷有 75 个项目。拟合检验结果显示有 51 题在 0.05 水平上无显著差异。

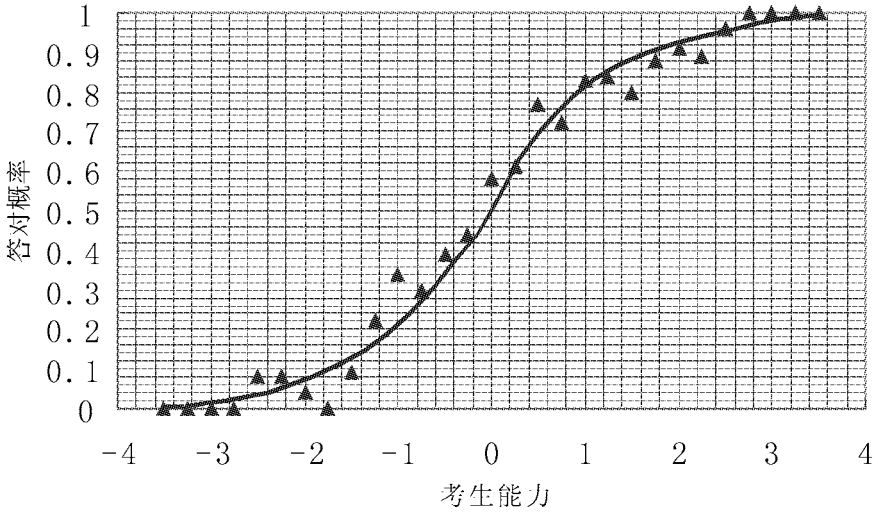

图 1　第 56 题的(a=0.49,b=－0.52)项目拟合图

该题为单等级(全或无)记分方式,此记分方式是等级反应模型的拟合检验的特例。图中实线为理论曲线,由 $P_j(\bar{\theta}_h)$ 计算得到;三角形图标为实际观察值,由$\frac{\sum_{k=0}^{m_j} r_{hjk}}{N_h}$计算得到。$G^2 = 6.93581$,自由度为 11,在 0.05 水平上检验无显著差异。

3.3　等值程序的校验情况

为了保证项目特征曲线法等值工作的科学性,我们对 PEB 系统的等值专用模块进行了校验。具体做法是使用由 15 个题(混合记分)组成的铆测验,对 4 种情况的等值常数,即等值常数分别为 $\alpha=1.5;\beta=0.5$,$\alpha=1.5;\beta=-0.3$,$\alpha=0.7;\beta=0.2$ 和 $\alpha=0.7;\beta=0.2$,利用蒙特卡洛方法,对每种等值常数,随机生成 20 批 x 和 y 测验上的项目参数,这样,共计有 80 批数据。模拟数据生成后,利用 Frank. B. Baker 编制的等值专用程序 Equate 2.1[7,8]与 PEB 里面的等值专用模块分别求出每批数据的等值常数的估计值 $\hat{\alpha}$ 与 $\hat{\beta}$,并求其跟等值常数 α,β 真值的绝对差的平均数和差值平方数的开方数,即 ABS* 与 RMSD**[10]值,看它们是否足够的小。显然,这两个指标是等值程序估计误差或者说功能强度的直接反映,自然是越小越好。

---

* $ABS: Abs = \sum_{t=1}^{n}(|\alpha_t^k - \alpha||\beta_t^k - \beta|)/n$

** $MSK: Rmsd = \sqrt{\sum_{t=1}^{n}((\alpha_t^k - \alpha)^2 + (\beta_t^k - \beta)^2)/n}$

表 2 等值程序的校验结果

等值常数	Equate 2.1				PEB			
	ABS		RMSD		ABS		RMSD	
	α	β	α	β	α	β	α	β
α = 1.5; β = 0.5	0.0412	0.0730	0.0588	0.0902	0.0169	0.0828	0.022	0.1097
α = 1.5; β = −0.3	0.0444	0.0673	0.0605	0.0857	0.0204	0.0816	0.0247	0.1005
α = 0.7; β = 0.2	0.0124	0.0189	0.01470	0.0235	0.0188	0.0281	0.0223	0.0352
α = 0.7; β = −0.2	0.0170	0.0297	0.02	0.0124	0.0099	0.0404	0.0394	0.0506

从表 2 结果看，我们自编的 PEB 中等值程序跟 Frank. B. Baker 编制的等值专用程序 Equate 2.1 的估计误差值的差异不是很大，一般来说只是在百分位的水平上略大（α 值还常略小），这意味着我们编制的等值程序其功能跟国外权威程序相仿，是可以实际使用的。

3.4 等值结果

因为人事部 2000 至 2003 年的经济师考试的实测数据（含初级、中级）基本上都可视为大体符合单维性假设，所以我们就采用项目特征曲线法把 2001 到 2003 年的试题参数全都等值转换到 2000 年的量纲系统上，实现了四个年度的测验等值。在等值设计中，虽然使用的是锚测验设计，但不是中心锚（即所有测验都使用相同锚题）设计，那么，我们就只能分别求两个相邻年度测验的等值常数估计值 $\hat{\alpha}$ 与 $\hat{\beta}$，即 2000 与 2001 年；2001 与 2002 年；2002 与 2003 年，分别求出其测验的等值常数估计值 $\hat{\alpha}$ 与 $\hat{\beta}$。就 2000 与 2001 这两个年度来说，具体操作是先使用 PEB 中的参数估计模块估计出每年测验的项目参数；然后，再分别将 2000 与 2001 年之间的锚题挑选出来形成两个参数文件，并将 2000 年锚题参数作为量表参数文件，2001 年锚题参数作为待转换参数文件，使用 PEB 等值模块求出 $\hat{\alpha}$ 与 $\hat{\beta}$；最后，用计算出的 $\hat{\alpha}$ 与 $\hat{\beta}$ 把 2001 年所有试题的项目参数转换到 2000 年量纲系统上。类似地，2001 和 2002 年度的数据，是先挑选出 2001 年与 2002 年锚题形成参数文件，并将 2001 年锚题参数作为量表参数文件（使用的参数是等值到 2000 年量尺上的参数），2002 年锚题参数作为待转换参数文件，然后用计算出的 $\hat{\alpha}$ 与 $\hat{\beta}$，把 2002 年参数转换到 2001 年的量纲系统上。因为 2001 年的项目参数已转换到 2000 年量纲系统上，这样做实际上是把 2002 年的试题参数转换到 2000 年量纲系统上。依此类推还可利用 2002 与 2003 年的实测数据求 $\hat{\alpha}$ 与 $\hat{\beta}$，并把 2003 年试题参数都转换到 2002 年（实际上是 2000 年）的量纲系统上。这样就将所有试题全部转换到 2000 年量纲系统上。经上述等值工作，我们转换了初、中级各 270 题，完成了经济学题库建设的关键步骤。

等值常数计算结果见表 3。

表 3 四年份测验等值常数计算结果

年份	初级		中级	
	α	β	α	β
2001 − 2000	0.84124	−0.26378	1.28863	0.17674
2002 − 2001	1.28978	−0.39367	1.58116	−0.20015
2003 − 2002	1.44895	−0.20303	1.72910	−0.26694

3.5 及格分数比较

等值技术的一个重要应用，就是以此来考察不同年份间合格标准确定的恰当性，进而指导合格标准的确定。利用等值转换后的结果，我们按人事部发布的 2000 年至 2003 年实际使用的及格分数，以 2000 年为标准，分析比较了这几年间及格水准的稳定性。

在项目反应理论中，式（2）形式的真分数，若除以测验的满分值，就成为掌握比例真分数，即有：

$$\pi_0 = \frac{1}{M_x}\sum_{i=1}^{n_v}\sum_{j=1}^{k_i} j_i \cdot p_{xvtij}(\theta_{xt}, a_{xvi}, b_{xvij}) \qquad (13)$$

这里，π_0 是掌握比例真分数而 M_x 是测验的满分值。当测验的所有项目参数均已知时，π_0 跟 θ_{xt} 是彼此一一对应的确定关系。正基于此，当几个年度的测验所有的项目参数都已等值到同一量纲系统上后，就可以直接比较各年度掌握比例是否处在同一水平上。具体操作是：先求得 2000 年的掌握比例分数（及格分数除以试卷总分），再计算掌握比例分数对应的能力值。然后，根据其他 3 年的项目参数计算当年得及格分数对应的能力值并比较其跟 2000 年的能力值是否相近。结果见表 4。

表 4　四年份测验的实际及格标准水平比较

年份	初级				中级			
	及格分数	总分	掌握比率分数	掌握分数对应的能力值	及格分数	总分	掌握比率分数	掌握分数对应的能力值
2000	56	130	0.43	0.1906	66	130	0.51	0.1472
2001	64	150	0.43	−0.1174	70	150	0.47	0.3984
2002	75	150	0.5	0.3794	75	150	0.5	0.2776
2003	70	150	0.47	−0.1357	76	150	0.51	−0.0092

表 4 说明,2000 年 2003 年连续 4 年的经济专业考试的合格标准中,从初级来看,2002 年有一个较大的波动(相对较高);从中级来看,2001 年和 2002 年的标准高于另外两年。如果要在这几年间都维持 2000 年的水准,即 $\theta_{初}=0.1906$,$\theta_{中}=0.1472$ 不变,借助式(13),可以反求它们在每个年度中的掌握比例分数与及格分数,见表 5。

表 5　及格标准保持不变时四年份测验应有的及格分数

年份	初级			中级		
	应有的能力值	对应的掌握比例	所对应的及格分数	应有的能力值	对应的掌握比例	所对应的及格分数
2000	0.1906	0.43	56	0.1472	0.51	66
2001	0.1906	0.46	70(69.7)	0.1472	0.45	67
2002	0.1906	0.48	72	0.1472	0.49	73
2003	0.1906	0.52	78	0.1472	0.53	79

从表 5 可以看到,每年考试为保持确能具有稳定的及格标准,可在考后先进行等值,再按上法求取与往年水平相同的本年度测验应有的及格分数值。2004 年考试部门已经使用这种方法来确定及格分数,并以此为参考向社会公布。当然,需说明的是,实际及格线的确定不能只考虑统计分析结果,还要考虑试卷的内容实际状况和社会用人需求等。

4　结论与讨论

本研究说明,人事部经济专业考试资料基本符合单维性假设,采用 Samejima 等级反应模型作分析是恰当的。在此前提下,运用锚测验等值设计,使用项目特征曲线等值法,自行开发编制专用等值程序计算模块,实现了 2000 至 2003 年 4 个年度考试资料的项目参数和能力参数等值,成功构建了经济专业题库。在此基础上,利用等值技术对不同年份试卷的及格分数进行了比较,为经济师考试合格标准制定、确保考试公平性提供了实证参考依据。

本研究中,锚测验题的选用存在一定不足,致使有的等值方程中的 α 系数值离 1.00 较远,有的接近或大于 1.5,这是今后应予改进的。另外,等值设计使用的是“链等值”或称“连环等值”方法,这种等值设计在等值次数多了以后,会累积一定的等值误差[9],如何减少误差,这也是今后应予以研究改进的。

参　考　文　献

1　Qi S Q, Dai H Q. Item response theory and its applications(in Chinese). Nanchang: Jiangxi Universities and Colleges Press, 1992
(漆书青,戴海崎. 项目反应理论及其应用研究. 南昌:江西高校出版社,1992)

2　Qi S Q, Dai H Q, Ding S L. Principles of modern educational and psychological measurement(in Chinese). Beijing: Higher Education Press, 2002
(漆书青,戴海崎,丁树良. 现代教育与心理测量学原理. 北京:高等教育出版社,2002)

3　Dai H Q. Study on test equating using method of Item Characteristic Curve transformation under Graded Response Model (in Chinese). Exploration of Psychology,2000,20(3):49 ~ 53
(戴海崎. 等级反应模型项目特征曲线法等值研究. 心理学探新,2000,20(3):49 ~ 53)

4　Zhu Z C, Yang H Z, Yang H R. Rasch model applied to score equating in the College English Test (in Chinese). Modern Foreign Languages, 2003,(1):69 ~ 75
(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):69 ~ 75)

5　Xie X Q. Comparison of 15 Equating Methods (in Chinese). Acta Psychologica Sinica, 2000,32(2):217 ~ 223)
(谢小庆. 对 15 种测验等值方法的比较研究. 心理学报,2000,32(2):217 ~ 223)

6　W J Van der Linden, Ronald K. Handbook of Modern Item Response Theory. Hambleton Eds, 1995

7 Kim S H, Cohen A S. A minimum method for Equating Tests under the Graded Response Model. Applied Psychological Measurement, 1995,19:167 ~ 176
8 Baker F B. Equating Tests under the Graded Response Model. Applied Psychological Measurement,1992,16:87 ~ 96
9 Kolen M J, Brennan R L. Test equating methods and practices. New York: Springer, 1995
10 Harwell M R. Analyzing the results of Monte Carlo Studies in Item Response Theory. Educational and Psychological Measurement, 1997,57(2):266 ~ 279.
11 Eiji M, Darreli B. Parscale Manual. Scientific Software Internation, Inc. 1997

Item Characteristic Curve Equating under Graded Response Models in IRT

Zhou Jun[1], Ou Dongming[2], Xu Shuyuan[1], Dai Haiqi[1], Qi Shuqing[1]
([1] *School of Education, Jiangxi Normal University, Nanchang* 330027, *China*)
([2] *National Personnel Examinations Authority, Beijing* 100080, *China*)

Abstract

In one of the largest qualificatory tests—economist test, to guarantee the comparability among different years, construct item bank and prepare for computerized adaptive testing, item characteristic curve equating and anchor test equating design under graded models in IRT are used, which have realized the item and ability parameter equating of test data in five years and succeeded in establishing an item bank. Based on it, cut scores of different years are compared by equating and provide demonstrational gist to constitute the eligibility standard of economist test.

Key words item characteristic curve equating, graded response models, item response theory