Rash 客观等距测量在 PISA 中国试测研究中的实践

王 蕾
(教育部考试中心 评价处,北京 100084)

摘 要:Rasch 测量是当前心理测量中具有客观等距量尺的测量,克服了传统经典测量的测验依赖和样本依赖的局限。以学生能力国际评价 PISA 中国试测研究为例,说明 PISA 如何应用 Rasch 测量达到跨越国家和地区教育成效比较的测量目的。客观等距量尺研究对改进和完善我国心理测量与教育评价具有重要参考价值。

关键词:PISA;教育评价;Rasch 模型;客观测量;等距量尺
中图分类号:B841.2 文献标识码:A 文章编号:1003 - 5184(2007)04 - 0069 - 05

心理测量,如认知、人格、态度、兴趣等的测量一直苦于无法达到物理测量般的客观和等距。乔治·拉什(Georg Rasch,1901 ~ 1980)创始的 Rasch 模型[1]成功的克服了这一困境。越来越多的学者认识到 Rasch 测量是可以使得心理测量达到客观等距的方法。在世界上具有较大影响的大规模国际教育评价项目 PISA(Programme for International Student Assessment),其阅读、数学和科学认知测验以及通过学生问卷和学校问卷收集的社会、文化、经济和教育因素的指标,就是通过 Rasch 测量进行题目的建构、参数校准和相关分析。PISA 如同用尺子量学生身高一样,用 Rasch 模型建构校准的,涵盖完整难易度的题目测量不同层次学生多维度能力发展水平、学习动机和态度等,进而进行国家间与地区间横向和跨年度纵向比较。这是国内普遍使用的依据经典测量理论建立的测评工具根本无法实现的。客观等距量尺研究对改进和完善我国心理测量与教育评价具有重要参考价值。

1 心理测量中顺序量尺当成等距量尺的错用

著名心理学家史蒂文斯(S.S.Stevens)1946 年提出了名义量尺、顺序量尺、等距量尺、比率量尺四种测量量尺,至今被广为采用。史蒂文斯还特别指出:“事实上,心理学家所经常使用的量尺是属于顺序量尺。严格来说,凡是会牵涉到平均数和标准差的统计方式,都不可以使用在顺序量尺上,因为这类的统计分析对量尺的要求,不单只是顺序就足够。”[2]心理测量中受试者认知测验的原始分、问卷调查的李克特量尺等只属于顺序量尺,却被错当成等距量尺,进行着平均数和标准差等统计推论。

顺序量尺使用上只有大小排序的意义。例如甲受试者的测验原始分 98 是班级第一名,乙受试者的测验原始分 96 是班级第二名,丙受试者测验原始分 94 是班级第三名。表面上 98、96、94 虽然有着 98 - 96 = 96 - 94 的关系,但本质上,很难说甲乙两生之间学业程度差异,恰等于乙丙两生的差异,只能说 98 > 96 > 94 的顺序关系是成立的。因此原始分只有顺序意义,无法精确描述其间的差距。在问卷调查里,常会要求受试者按李克特量尺来做答。以五点李克特量尺为例:“对于部分高校实行 5% 名额的自主招生,您的意见是 1 完全同意、2 比较同意 、3 一般、4 比较不同意、5 完全不同意”。研究者常分别以 1,2,3,4,5 代表这五点。这五点只能算是顺序量尺,因为没有证据显示 1 到 2 的差距等于 2 到 3 的差距。将这些题目的得分加总后的总分也只是顺序量尺。

等距量尺不仅有顺序意义,还有差距的意义。例如摄氏 25 度和 26 度之间的差距,等于 26 度和 27 度之间的差距。简而言之,1 度的差距在任何度数上都有相等的距离,因此称为等距量尺。

2 Rasch 客观等距测量的分析程序

在传统的测验里用受试者的得分来定义其程度,用答对率来定义题目的难度。如果测验很简单,受试者的得分就高,显示受试者程度很好。反之,如果测验很难,受试者的得分就低,显示受试者程度很差。到底受试者的程度是好是差,取决于测验的特性,因此是测验依赖。同理,当判断题目难度时,如果受试者的程度很差,答对率就低,显示该题难度很高。反之,如果受试者程度很好,答对率就高,显示该题目难度很低。到底题目是难是易,取决于受试者样本的特性,因此是样本依赖。总之,受试者的能

力估计和题目的难度估计是彼此干扰,没有‘客观等距’可言。当测验依赖和样本依赖无法有效解决时,所得到的测验分数在使用和后续分析都会违背测量的目的和统计假设而存在问题。如何积极寻找和使用等距量尺达到测量的目的,有效解决原始分分析使用中没有客观等距的实际问题。

乔治·拉什 1960 年提出了 Rasch 模型,就是希望透过受试者的作答反应,得到客观等距的量尺。Rasch 模型测试的潜在特质其实就是测验工具想要测量的目的,它可以是能力、也可以是人格特质、态度、兴趣、价值观等。

当分析的真实数据拟合 Rasch 模型的预期所得到的量尺就有很好的测量特性。为得到拟合 Rasch 模型的数据首先要利用已有的作答反应数据通过最大似然估计法估计最有可能的受试者的潜在特质和题目的难度。然后判定数据拟合 Rasch 模型的拟合度。例如 PISA2006 中国试测每道题目约有 1500 位受试者作答,因此就有了约 1500 个观察值。透过 Rasch 模型的分析软件可以计算出相应这 1500 个期望值,然后计算实测数据与期望值得残差,判定这些残差是否大得离谱。假如很多能力低的人答对该题,但很多能力高的人却答错,就判定该题不拟合 Rasch 模型预期。这是题目拟合度(item – fit)分析。将不拟合 Rasch 模型预期的题目删除。常规的题目分析程序还包括项目功能差异(differential item functioning;DIF)检验,也就是要判定某个题目对不同群体而言是否都具有相同意涵。DIF 题目指尽管受试者能力相当,由于他们来自不同群体,如男、女学生或不同国家受试者,答对该题的概率却会不等。DIF 题目对不同群体产生了不公平的测量,此类题目应从测验分析中删除从而确保所有题目对每个群体的受试者都在测量同一潜在特质,且其分数可以比较。

3 Rasch 测量在 PISA 中国试测研究中的实践

3.1 PISA 简介

学生能力国际评价 PISA 是经济合作与发展组织(Organization for Economic Co – operation and Development;OECD)发起并组织实施的为各参与国家与地区协作监控教育成效的评价项目。测试发达国家与地区义务教育阶段结束后 15 岁的学生在阅读、数学和科学领域的发展水平。PISA 通过收集学生的背景信息,进行包括个人、家庭和学校等方面的因素解构,形成教育成效评价指标体系,为各参与国家与地区政策分析和研究提供有价值的参考。PISA 在 2000 年首次开始评价,每三年进行一次,周而复始,以评价年命名。PISA2000 有 32 个国家参与,PISA 2003 有 41 个国家和地区参与,PISA2006 有 56 个国家和地区参与[3]。为了保证评价的效度和信度,由来自各参与国家和地区的教育政策制定者和相关领域的专家共同决定评价的范围、本质、目的等。评价材料也考虑到不同的文化和语言影响,其翻译、取样和资料收集过程都采取了严格的质量监控机制,并通过实施大规模实地预试等各种手段,将测试在各参与国家与地区实施中可能存在的误差降到最小。

PISA 为各参与国提供了精准的测试工具测量义务教育结束阶段 15 岁学生科学、数学、阅读素养。PISA 测量的素养[4]是指 15 岁在校学生,为迎接当今不断变化的现实世界的挑战,应用知识和技能解决问题的能力,以及在日常生活情境下做出良好判断和决策的能力。它不同于且高于对于学校课程所设置的学科相关知识的理解或记忆能力。PISA 测试工具测量的是 15 岁学生的认知整体发展水平,涉及彼此关联的阅读、数学、科学素养,而非传统意义上的学生学科成绩。通过严格抽样的学生样本认知发展水平的数据分析,推断各参与国家与地区的教育成效。为实现这一测量目标,必须使用具有客观等距量尺的 Rasch 测量,达到跨地区横向和跨年度纵向比较。

PISA 试题提取各学科课程的内在联系,结合学科的特点,设计了有深厚学科理论背景的问题,要求学生采用应用和探索的方法,在对学科知识融会贯通的基础上解决实际问题。PISA2006 经过在 56 个参与国家与地区大规模实地预试,排除不同文化背景和语言翻译的影响,精心选用了题库中 Rasch 测量属性良好的 28 道阅读题,48 道数学题和 108 道科学题目完成了 15 岁学生阅读、数学、科学素养测试,并进行了测试结果的国际比较与趋势分析。

3.2 PISA 中国试测研究

教育部考试中心 2006 年 10 月启动了 PISA 中国试测研究。PISA 中国试测研究并不代表我国正式参与 PISA,目的在于通过实践,掌握、借鉴 PISA 先进的评价理念、理论、技术,构建符合中国国情的评价标准、手段、技术和方法体系;促进考试内容和形式的改革,有利于全面推进素质教育。PISA 中国试测研究按 PISA 国际规范采用两阶段分层随机抽样设计,从试点地区抽取了 150 所样本学校,随后以完全随机抽样的原则从这 150 所学校样本中抽出 5000

余名学生作为参加测试的学生样本。样本有效地代表了试点地区近 1200 所学校的 16 万名 15 岁在学生总体，其中农村学校在校学生将近一半。

PISA 中国试测研究采用的是纸笔测验，共有 13 套试题册，依据随机原则将每名学生样本分配到每套试题册，每个学生需用两个小时的时间完成测验，随后作答一份约半小时的学生问卷。学校样本的校长要求作答一份约 20 分钟的学校问卷。PISA 中国试测研究收集的原始数据同时采用国际大型统计软件 SPSS 和 SAS 两套系统进行数据清理和转换，得到完全匹配的结果，保证了数据的精准。随后，用清理后的数据按照 OECD 数据分析的标准流程使用 Rasch 模型分析软件 ConQuest[5] 对每个学生样本依据所属不同学校、所做不同试题册、学生问卷收集到的不同背景类型进行回归分析，每个认知测量领域生成 5 个似真值（plausible value，PVs）[6]，对 PVs 进行加权，通过与抽样相关的基于 SPSS 开发的多重复制程序（replicate）[7] 进行分析，按国际规范得到了 Rasch 测量客观等距量尺的测量结果。因为 PISA 中国试测数据与 Rasch 模型的预期拟合良好，测量结果可直接与参与 PISA 的 56 国家和地区相比较。

3.3　PISA 中国试测研究试题拟合度和项目功能差异研究

PISA 为确保国际比较结果的精准性，所有题目除进行 Rasch 模型拟合度分析外，还进行全方位 DIF 分析。在 PISA2006 通过预试最终选定的 108 道科学素养题目正式测试后进行结果分析时，仍发现有 5 道题目与 Rasch 模型预期相差甚远或存在 DIF，故国际数据校准时删除了这 5 道题目的数据后才进行国际比较。

Rasch 模型分析软件 ConQuest 提供各类数字、图表指标进行题目拟合度和项目功能差异分析。PISA 中国测试研究以检验平均标准误差（infit MNSQ）数值是否超出 0.75 到 1.33 的临界范围作为检测标准。经 4892 个有效学生样本在数学、阅读、科学领域题目拟合度分析表明中国试测地区学生对所有题目的总体反应与 Rasch 模型的预测相当拟合。图 1 显示了 PISA2006 一道数学题与中国测试学生实测数据的题目拟合度特征曲线图。图中实线为 Rasch 模型的预期，虚线为中国试测学生作答反应。从图中可以看出此题目拟合度良好。

用 ConQuest 软件进行中国测试研究试点地区 DIF 分析。题目显示实质性 DIF 检测标准为试点地区间难度差异统计在 0.05 水平显著并且 logit 之差大于 0.7。依据此标准，阅读只有 1 题、数学 6 道题，科学 8 道题显示试点地区间实质性 DIF。PISA 中国试测数据在进行分地区、性别、社会经济地位等不同群体教育成效比较时，删除了实质性 DIF 的题目，得到客观等距的测量结果后才进行了有效的统计分析。

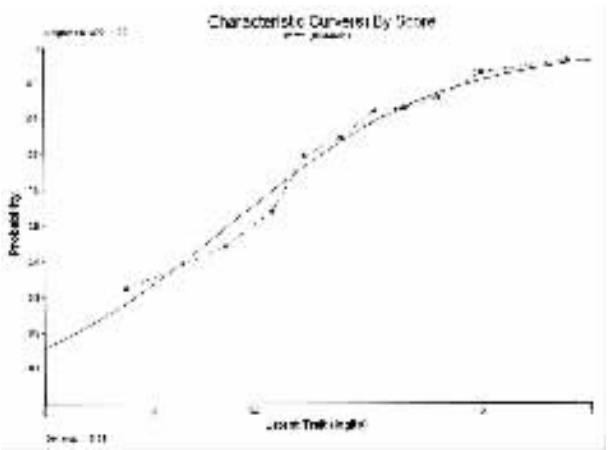

图 1　PISA2006 中国试测研究题目拟合举例

PISA 利用 Rasch 模型和先进的数据分析方法，在时间短、学生样本量小、题目样本量大，覆盖面广、差小、减负好的同时，确保了测量的科学性，拓宽了评价的内容与形式，使教育评价更具实用价值。

3.4　PISA 客观等距量尺的国际比较

PISA 从 30 个 OECD 国家的每个参与国随机挑选出 500 个参与测试的学生样本用作国际参数校准的有效样本。题目国际参数在这 15000 个随机挑选的学生样本的基础上通过阅读、数学、科学单维校准获得[8]。

通过 ConQuest 软件锚定 PISA2006 国际参数，采用多维随机系数多项逻辑斯蒂模型（Multidimensional Random Coefficients Multinomial Logit Model）[9] 可标定出各参与国家或地区学生样本所代表的学生素养发展水平分布。学生不同素养的测量是相互依赖的。一个学生在理解科学术语的时候，阅读素养是必要的，解释数据时也需要数学素养。为能得到更精确的学生素养测量结果，图 2 为锚定 PISA2006 国际参数后，中国学生样本在数学、阅读、识别科学问题、科学地解释现象和使用科学证据五个维度同时校准的怀特图（Wright Map）。此图显示了学生样本和考题在对应的能力水平定义范围内的展开分布。左边每个 X 代表 20.8 个学生样本，共有 4892 个有效学生样本，右边为题目编号。

	数学	阅读	科学地解释现象	识别科学问题	使用科学证据	题目
	XXXX	X	X	XXXX	X	
	X					
	X					
	X					
	X					
	X					
	X					
3	XX					25
	XX		X			
	XX	X	X		X	
	XX	X	X	X	X	
	XXX	X	XX	X	X	
	XXXX	X	XX	X	X	
	XXXX	X	XX	X	X	
	XXXX	XX	XXX	XX	XX	
	XXXXX	XX	XXX	XX	XXX	
	XXXXX	XXX	XXXX	XX	XXX	61
	XXXXXX	XXX	XXXX	XX	XXXX	133 177
2	XXXXXX	XXXX	XXXXX	XXX	XXXX	
	XXXXXX	XXXXX	XXXXXX	XXX	XXXXX	14 20 27 79
	XXXXXX	XXXXX	XXXXXX	XXXX	XXXXXX	
	XXXXXX	XXXXXX	XXXXXXX	XXXXX	XXXXXX	98
	XXXXXX	XXXXXX	XXXXXXXX	XXXXX	XXXXXXX	56
	XXXXXXX	XXXXXXX	XXXXXXXX	XXXXXX	XXXXXXXX	44
	XXXXXX	XXXXXXX	XXXXXXXXX	XXXXXX	XXXXXXXX	5
	XXXXXX	XXXXXXXX	XXXXXXXXX	XXXXXX	XXXXXXXXX	60 114 172
	XXXXXXX	XXXXXXXX	XXXXXXXXX	XXXXXXX	XXXXXXXXX	28 63 100
	XXXXXXX	XXXXXXXXX	XXXXXXXXX	XXXXXXX	XXXXXXXXX	13 81
	XXXXXXX	XXXXXXXXX	XXXXXXXXX	XXXXXXXXX	XXXXXXXXX	41 47 48 176
1	XXXXXXX	XXXXXXXXXX	XXXXXXXXX	XXXXXXXX	XXXXXXXXX	11 57 68 86 132
	XXXXXXXXXXXXXXXXXXXXXXXXXX			XXXXXXXX	XXXXXXXXX	78 90 137 152 170
	XXXXXXX	XXXXXXXXX	XXXXXXXXX	XXXXXXXX	XXXXXXXXX	21 38 39 64 105 110 158 166
	XXXXXXX	XXXXXXXXX	XXXXXXXX	XXXXXXXX	XXXXXXXXX	93 102 113 128 161
	XXXXXX	XXXXXXXXX	XXXXXXXX	XXXXXXXX	XXXXXXXXX	45 89 103 129 155
	XXXXXX	XXXXXXXXX	XXXXXXX	XXXXXXXX	XXXXXXX	23 50 109 112 131 147 156
	XXXXXX	XXXXXXXXX	XXXXXXX	XXXXXXXX	XXXXXXX	82 91 121 160 168
	XXXXXX	XXXXXXXX	XXXXXXX	XXXXXXXX	XXXXXXX	7 54 66 123
	XXXXXX	XXXXXXX	XXXXXX	XXXXXXXX	XXXXXXX	2 17 73 75 80 101 169 171
	XXXXXX	XXXXXXX	XXXXXX	XXXXXXX	XXXXXX	15 33 34 37 70 74 135 153
	XXXXX	XXXXXX	XXXXXX	XXXXXXX	XXXXXX	35 69 87 94 148 159 165
0	XXXXX	XXXXXX	XXXXX	XXXXXXX	XXXXX	30 51 77 108 162 179
	XXXXX	XXXXXX	XXXX	XXXXXX	XXXXX	16 18 55 95 124 139 173
	XXXX	XXXXX	XXXX	XXXXXXX	XXXXX	72 97 134 146 150 157 178
	XXXX	XXXX	XXXX	XXXXXX	XXXX	8 46 65 88 99 130 136 143
	XXX	XXXX	XXXX	XXXXXX	XXXX	6 62 96 111 163 175
	XXX	XXXX	XXX	XXXXX	XXXX	4 92 118 119
	XXX	XXX	XXX	XXXXX	XXX	12
	XXX	XXX	XX	XXXX	XXX	36 76 104 115 127
	XXX	XX	XX	XXXX	XXX	3 19 32 42 107 149 151
	XX	XX	XX	XXXX	XX	31 52 85 116 141 142
– 1	XX	XX	XX	XXX	XX	24 122 138 144
	XX	XX	X	XXX	X	120 145
	XX	XX	X	XXX	X	26 43 67 125 164
	X	X	X	XX	X	71 106 117 140
	X	X	X	XX	X	29
	X	X	X	XX	X	58 59 83
	X	X	X	XX	X	1 49 154
	X	X		X	X	
	X			X		174
	XX			X		9 10 22 40 53 84 126 167

图 2 锚定 PISA2006 国际参数中国试测学生五个维度 Rasch 模型同时校准的怀特图

上图显示出 PISA 如同测量学生身高一样，用通过 Rasch 测量建构的试题打造了一把测量学生素养的精准量尺。PISA 通过 Rasch 测量打造的是一把钢性的量尺，试题难度不会象传统测试的弹性量尺随受试人群样本的能力不同而变化。各参与国家与地区一致认同 PISA 客观等距量尺测量出的教育成效，其高品质保障的取样、测试管理机制和最新的数据后期分析使 PISA 跨地区和跨年度比较具有高度的信度和效度，同时又如同比较学生身高一样简单明了。

PISA 通过 Rasch 测量所提供的具备等距量尺特性的测量结果可以克服传统测验原始分数据分析的缺陷。传统测验原始分数据分析误将顺序量尺题目的加总分数视为具备等距量尺的特性，并进而使用相应统计分析推论到总体，不能真实反映出总体情况，导致错误结论。PISA 拥有 5000 多道由各参与国家或地区贡献，并经过 Rasch 模型拟合检定的题库。PISA2006 通过选用题库中的题目组成 13 套试题册，Rasch 测量分析不但实现了不同试题册的等值，还与 PISA2000 和 PISA2003 的测验链接，实现跨年度的纵向比较。

Rasch 测量的主要目的在于解决心理测量上的困境，透过 Rasch 模型来检验测验的数据。如果测验数据拟合 Rasch 模型的预期，就可以得到客观等距量尺。如果数据不拟合模型的预期，例如某受试者专门答对很难的题目，却答错很简单的题目，亦或某道题目专门被能力低的受试者答对，而被能力高的受试者答错，那么此测量应宣告失败。

Rasch 测量已经为心理测量和教育评价带来革命性的影响，从 1960 年 Rasch 测量创始到现在 40 余年来，经过多位学者的努力，使得 Rasch 测量有相当多元的发展，形成了 Rasch 家族模型，可以适用于多分题、多相、多层次、多维度、潜在类别等复杂的测验情境。Rasch 客观等距测量量尺研究无疑对改进和完善我国心理测量与教育评价具有重要参考价值。

参考文献

1　Georg Rasch. Probabilistic models for some intelligence and attainment tests. Copenhagen: Institute of Educational Research, 1960.

2　Stevens S S. On the theory of scales of measurement. Science, 1946, (103): 667 – 680.

3　OECD. Assessing scientific, reading and mathematical literacy: A framework for PISA 2006. Paris: OECD, 2006.8.

4　OECD. Learning for tomorrow's world: First results from PISA2003. Paris: OECD, 2004.25.

5　Wu M, Adams R J, Wilson M. ConQuest. Hawthorn, Australia, ACER Press, 1998.

6　7　OECD. PISA2003 Data Analysis Manual. Paris: OECD, 2005.31 – 80.

8　9　OECD. PISA2003 Technical Report. Paris: OECD, 2005.119 – 134.

Implementation of Objective Measurement with Interval Scale in Psychology Measurement into PISA China Trial

Wang Lei

(National Education Examinations Authority, Ministry of Education, Beijng 100084)

Abstract: The Rasch model is the model that yields objective measures with interval scale in current psychology measurement. Taking PISA China trial test data as an example, this paper demonstrates cross – country comparison on student performances is possible only when data meet the expectation of the Rasch model. Valuable references have been gained to improve current psychological measurement and educational assessment in China.

Key words: PISA; educational assessment; Rasch Model; objective measurement; interval scale