心理学报 2009, Vol. 41, No.8, 773−784
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2009.00773

GT 与 IRT 的比较: 北京奥运会男子 10 米跳台跳水分析

俞宗火[1] 唐小娟[2] 王登峰[1]

([1]北京大学心理学系, 北京 100871) ([2]南昌航空大学数信学院, 南昌 330063)

摘 要 概化理论(GT)和项目反应理论(IRT)从两个不同的方向发展了经典测量理论, GT 和 IRT 中的多面 Rasch 测量模型(MFRM)在主观评分中都可以用来估计评分中各变异来源对变异的贡献, 对测评的信度进行估计, 提出测评改进意见。12 名运动员参加了 2008 北京奥运会男子 10 米跳台跳水决赛, 比赛共 6 个回合, 7 名裁判独立对他们在各个回合的表现进行打分。GT 和 MFRM 比较一致地认为运动员自身、回合、运动员与回合的交互效应是运动员得分的重要变异来源, 而裁判员对运动员得分差异的贡献不显著。MFRM 同时还估计出难度系数是影响男子跳台跳水成绩的重要变异来源, 在评分等级 6.5 附近存在步校准错乱, 得出的运动员成绩排序与 2008 奥运实际排序有所不同。在 GT 中难度系数作为隐藏侧面, 其效应未能分离出来。GT 和 MFRM 从两个不同的方面给测量提供改进意见: GT 发现可以通过增加回合数来提高 g 系数, 而增加裁判数对其影响不大。MFRM 给出各侧面的要素(如某裁判、运动员等)的估计值及其标准误, 它给出的诊断性拟合统计也有助于甄别异常得分或评分模式。

关键词 概化理论; 多面 Rasch 测量模型; 主观评分

分类号 B841;B849

1 引言

测量是按照一定规则给研究对象在一定性质的数字系统(尺度)上指定数值的过程, 其目的在于正确地认识和对待客观对象。由于心理现象有非实体性、动力性等特点, 心理测量所直接测到的只能是作为心理活动过程和心理特性作用结果的外显行为(漆树青, 戴海崎, 丁树良, 2002)。心理测验就是在标准化情景下对一组行为样本的测量。我们真正关心的是一般情况下的更具一般性的行为, 而非特定时空条件下的表现。与行为抽样相关联的一个问题是, 个体在特定时空条件下的行为表现是否可以推论到一般情况下的更具一般性的行为, 这种推论有多大的把握。而事实上, 标准化情景并不能保证不同时空条件下的测验是真正的“平行测验”。换句话说, 不同时空条件下的测量, 被试受情景等的影响, 其本身的行为表现可能是不同的。其中主观评分的测验则更为复杂, 被试的得分不仅受到特定时空条件下被试自身行为表现的影响, 还会受到来自评分者的影响。在主观评分活动中, 被测者按照规定完成一定的综合任务, 评分者则根据自己的主观判断对于被测者完成任务的具体表现给出一个综合分数。评分者的评分容易受到其知识水平、综合能力、爱好、情绪、疲劳等主观因素的影响。因此, 主观评分活动中, 可能会出现同一个评分者前后宽严标准不一、不同评分者宽严标准不一的情况, 出现所谓的评分者效应(Wolfe, 2004) 。研究者和测验管理者的目标就是要减小这些误差, 准确估计被试的能力值。

以作文能力的测量为例, 根据经典测量理论, 要减少测量误差, 准确估计被试的写作水平的能力值, 一个办法就是让考生写不同体裁、不同主题的论文, 每样写几篇, 同时每篇请不同的评分者打分, 然后取平均值, 就可以得到该考生作文能力的真分

收稿日期：2008−11−14

通讯作者：王登峰, E-mail：wdfeng@pku.edu.cn

数。但是, 这种思路既不现实, 也不可行, 因为让考生写那么多的文章为时间精力等所不容许, 同时, 究竟需要写多少篇论文, 要多少个评分者来评分, 仍然是个未知数。为了克服经典测量理论的局限, 现代测量理论中的两个非常重要的理论(概化理论和项目反应理论)发展起来了。但这两者走的是不同的发展思路, 概化理论与经典真分数理论一样, 属于随机抽样理论, 它从测量的外部或宏观方面入手, 重在讨论测量的外部效度问题, 采用统计学中的方差分量模型, 具体分析实际的测量情景关系, 根据不同的确定测量目标和侧面的做法, 有针对性地考察多种信度和效度。项目反应理论则是一种量表化模型理论, 从测量的内部或微观方面入手, 着重讨论测量的内部效度问题, 考察被试的特质水平与测验项目之间的实质性关系(漆树青, 戴海崎, 丁树良, 2002; 杨志明, 张雷, 2003)。可以认为, 概化理论更多地涉及到减少误差的问题, 通过概化理论估出各个侧面的变异大小, 并通过研究设计, 提出减小误差的测量设计; 而项目反应理论更多的涉及到如何准确估计被测者特质水平的问题。两个理论承认观察分数与真分数不完全等同, GT 能告诉我们的是, 观察分数与真实分数之间的关联究竟有多大, 它的输出是概化系数, 这个信度只能概化到类似的情景中去, 因为观测值只是既定全域的一个抽样; 而 IRT 则是剔除其他因素的影响, 对观察分数进行校正, 使之尽可能地接近真分数, 输出的是真分数的估计值, 这个估计值可以推论到性质相同但量上有差异的情景(如更严的裁判等)中去(Cronbach, Gleser, Nanda, & Rajaratnam,1972; Linacre, 1993)。同时, GT 和 IRT 的立足点是不同的, GT 强调项目、裁判的可替换性及同质性, 而 MFRM 鼓励的是项目之间的差异性及裁判自身的一致性(Everett & Kulikowich, 2004) 。当主要任务是要考察当前情景下测得的目标群体的原始分数与其他类似情况下测得的原始分数的相似性时, 应该选用 GT; 而如果是要考察剔除其他因素的影响, 各考生的真分数是多少时, 应该选用 MFRM (Linacre, 1993)。

四年一度的奥运会举世瞩目, 不仅是当今世界影响最大的体育比赛, 更是各国交流文化、展现国力的盛事。一个国家获取奖牌数量的多少, 也是各界奥运会中备受瞩目的一件事。每一场比赛, 不仅仅是各国运动员展现实力与风采的机会, 实际上也是一个测量的过程, 运动员的成绩不仅受运动员自身技能水平的影响, 也受到比赛情景的影响。此外, 很多比赛项目中, 涉及裁判的主观评分, 裁判员要在比赛进行的过程中迅速做出裁判, 裁判的情况成为影响运动员成绩的一个方面, 甚至是引起争执的一个方面。介于现代测量理论的发展, 本研究跳台跳水项目为例, 试图引入概化理论和项目反应理论同时对 2008 北京奥运会十米男子跳台跳水的评分进行分析比较, 一方面我们要考察在当前的比赛及评分规则下, 目标群体的原始分数与真分数之间的关联究竟有多高, 如何能提高两者之间的关联性; 另一方面, 考察其他相关因素对评分的影响模式, 尽可能给每个参赛者的能力值一个精准的估计。为进一步提高奥运会主观评分项目的科学性提供思路。

1.1 概化理论

在经典测量理论中, 误差和信度都没有一个固定不变的定义。比如, 将信度视为稳定性的估计时, 误差是由于在不同时间测量引起的; 将信度视为内部一致性时, 误差则是由于不良的行为抽样引起的。因此, 一个基于经典测量理论的信度系数, 只能指向一个测量误差来源。当需要报告多个信度系数时, 可能会出现大小不一甚至相差很大的情况。这时, 决策者可能会无法决定该如何使用这些信息。概化理论的一个优点就是能够整合不同的测量误差来源及它们之间交互效应的信息。在概化理论中, 观察分数被认为是从观察全域中抽样出的一个样本, 观察分数的变异分解为测量目标的方差分量(相当于经典测量理论中真分数所贡献的变异)、其他测量侧面的方差分量、测量目标与测量侧面交互效应的方差分量及测量侧面之间交互效应的方差分量(此三者相当于经典测量理论中, 测量误差所贡献的变异)。方差分量是针对单个观察而言的, 根据方差分量的大小, 我们就可以知道各侧面变异来源的大小。概化理论根据这些信息构造出一个相当于经典测量理论中的信度相类似的概念—概化系数(Brennan, 2001a; 漆树青 等, 2002)。

概化理论研究分为两步, 第一步为概化研究, 又称为 G 研究。G 研究的主要任务包括: 明确测量的对象和测量的目标, 确定测量的侧面和观测全域以及测量设计和测量模式, 最后是用实验设计的思想方法来分解总变异, 将测量目标以及众多的测量侧面的效应及交互效应估计出来。作为 G 研究输出的产品, 主要包括下面四个信息:

1) 绝对误差。除了测量目标引起的变异之外,

其他所有方差分量之和就是绝对误差。

2) 相对误差。所有测量目标与测量侧面之间的交互效应的方差分量之和就是相对误差。

3) 概化系数。又称 g 系数, 测量目标方差除以它本身与相对误差的和, 所得到的商, 就是概化系数。概化系数的取值范围为 0~1.0。

4) 可靠性指数。又称 phi 系数, 测量目标方差除以它本身与绝对误差的和, 所得到的商, 就是可靠性指数。其取值范围为 0~1.0。

概化研究的第二步为决策研究, 又称为 D 研究。D 研究的任务是根据 G 研究的结果, 通过调整全域中各个侧面的样本容量、测量的模式及测量结构, 重新构建多种概括全域, 在样本均值的层面上估计各种变异分量的大小, 进而估计各种测量误差和测量精度指标, 从而从中选取较为合理的测验设计, 为决策提供依据(Brennan, 2001a; 杨志明等, 2003)。

1.2 多面 Rasch 测量

在经典测量理论中, 项目分析是依赖于被测者的抽样的, 而对被试能力的估计又依赖于特定的测验项目, 能力量表和难度量表不是定义在同一个尺度之上, 从而无法判断被试对某个题目会做出怎样的反应, 影响测量的选题。此外, 经典测量理论和概化理论都假定测验得分是等距量表, 而实际上, 无论是 Liker 量表还是成就测验的得分, 都是等级量表。经典测量理论的这些缺点, 为项目反应理论所克服(Embretson & Reise 2000; 漆树青 等, 2002)。

多面 Rasch 测量模型(MFRM)是对项目反应理论模型中的基础 Rasch 模型的拓展。MFRM 不仅和其他项目反应理论模型一样能克服经典测量理论的上述缺点, 而且能够估计评分者之间的不一致、测量情景的不一致、任务难度的不一致等引起的影响。MFRM 会给出各个侧面的每一个体的估计值, 对于各个侧面上个体的得分与这个估计值之间的匹配程度, MFRM 会提供“拟合统计”, 这就是“infit”和“outfit”均方值。在 MFRM 模型中, 这两个统计量的期望值是 1.0, 标准差为 0, 其合理的取值范围, 大多认为应该在 0.5~1.5 之间, Michael Linacre 认为小于 1.5 是比较好的, 在 1.5~2.0 之间则不好不坏, 大于 2.0 则被认为不好(Sudweeks, Reeve, & Bradshaw, 2005)。但也有很多研究认为应该在 0.8~1.2 之间比较合理(Engelhard, 1994a, 1994b; Linacre & Engelhard, 1994; 孙晓敏, 张厚粲, 2006)。

此外, MFRM 模型能够分析各侧面之间的交互效应, 这种交互效应又称为侧面功能差异(differential facet functioning)。如某个裁判员倾向于给某类被测者打高分, 这就是裁判员功能差异; 某个题目更有利于某个特定人群得高分, 这就是项目功能差异。MFRM 模型给侧面间的每个交互效应估计出一个以 logit 为单位的“偏差(bias)”分及相应的一个 z 分数, 如果 z 分数的绝对值大于等于 2, 就说明交互效应显著 (Linacre, 2003)。

1.3 概念比较

在比较概化理论和项目反应理论的 MFRM 时, 我们有必要对两者都使用但在界定上又有一些微妙的差异的几个非常重要概念进行澄清。这些概念包括侧面(facet)、交互效应(interaction)和信度(reliability)(Sudweeks, et al, 2005)。

在概化理论中, 总是希望测量目标的变异越大越好, 而测量目标以外的测量侧面则被认为是系统误差的来源; 而 MFRM 则认为所有的侧面都是模型的系统的变异来源。如本研究中的运动员、回合、裁判员, 在概化理论中, 运动员被认为是测量的目标, 回合和裁判员是测量侧面, 而在 MFRM 中, 三者都是侧面。

概化理论中交互效应与方差的析因分析中的交互效应很类似, 对于每一个主效应间的交互效应, 只报告一个方差分量的相对大小。而在 MFRM 中, 各侧面间的交互效应被认为是侧面功能差异, 如项目功能差异等等, 以一个个案例偏差的形式给出报告。比如, 本研究中, 如果运动员与裁判员之间交互效应显著, 在概化理论来说, 它只给出一个总的“运动员×裁判员”方差分量的估计, 而在 MFRM, 所有的运动员与所有的裁判员之间, 一一都有一个交互效应的估计。

概化理论中的 g 系数和 phi 系数, 都被认为是领域抽样模型中的信度系数。它们用来说明单个考生的得分的均值可以被用来预测全域分数的程度。在 MFRM 的分析报告中, 对每个侧面都有两个不同的“信度”: 分离信度(separation reliability)和分离比(separation ratio)。这两个统计量都是用来描述测量中 MFRM 估计的特定侧面的不同要素间的变异大小的指标, 但这两者所使用的尺度不同: 分离比是 Adj. SD 除以 RMSE 得到的商, 其变动范围为 0 到无穷大。分离信度则是真实变异占总观测变异的比例的量度, 其变动范围在 0 和 1.0 之间。对于不同的侧面来说, 这两个统计量的大小的意义是不同的。一般来说, 对于被测量者这个侧面来说, 高分

离信度和分离比是好的, 而对于其他研究者不感兴趣的侧面来说, 一般总是希望这两个统计量比较小才好。因为其他侧面的要素间的变异, 所代表的是与测量的构念无关的变异。而被测量者侧面, 则相当于 Cronbach 的α系数, 值越高表示测量越可信(Sudweeks, et al, 2005) 。

2 方法

2.1 被试

12 名来自世界各地参加 2008 年北京奥运会十米男子跳台跳水决赛的运动员。就本届奥运会而言, 这 12 名运动员便是该项目上决赛运动员的全体, 同时, 我们可以认为前后几届奥运会决赛运动员的总体水平是相当的, 从这个角度来说, 这 12 名运动员也可视为该项目上从前后几届奥运会决赛运动员的一个抽样。

2.2 程序

2008 年奥运会跳台跳水决赛共有六个回合比赛。根据运动员起跳前站立的方向和起跳后身体运动的方向, 跳水动作分为向前、向后、反身、向内、转体、臂立 6 组。在完成动作的过程中, 运动员可采用直体、任意姿势、屈体或抱膝等姿势。运动员在不同的回合里的动作必须在不同组别中选出, 不能重复。各运动员可自行决定表演这六组动作的顺序。每个跳水动作一般有 3 或 4 位数外加一个字母来标示*, 以表示动作组别和翻腾转体的周数等。比赛共有 7 名裁判, 分别独立对运动员的成绩进行评定。每个回合的比赛最高评分为 10 分, 最低 0 分, 以 0.5 进位。失败 0 分, 不好 0.5~2 分, 普通 2.5~4.5 分,较好 5~6 分,很好 6.5~8 分,最好 8~10 分。裁判员根据运动员的助跑(即走板、跑台)、起跳、空中动作和入水动作来评定分数。比赛计分规则是 7 名裁判员打出分数以后, 先删去两个最高分和两个最低分, 余下 3 名裁判员的分数之和乘以运动员所跳动作的难度系数, 便得出该动作的实得分, 六个回合的得分加总便是总成绩, 然后根据总成绩排序。本研究中使用的运动员的得分是未删减、未按难度系数加权的原始数据。

2.3 数据分析

从动作的编码, 我们也容易知道, 即便来自同一组的动作也是千变万化的, 故, 本研究中不采用动作的组别来作为变异的来源, 只考虑难度系数、回合和运动员和裁判员四个变异来源。其中动作难度系数是表明运动员完成动作的难易程度。动作简单, 难度系数就低; 动作复杂, 难度系数就高。国际跳水竞赛规则为每一个跳水动作确定了相应的难度系数, 它根据动作组别、竞赛项目(跳板、跳台)、器械高度、动作姿势和翻腾转体的周数等方面的差异来确定其数值。所以, 难度系数的差异能够比较好地表明动作本身的差异。

采用 Robert L. Brennan 编写的 urGENOVA 2.1 软件进行概化理论的分析。在实测数据中, 各位运动员在一个回合中只能选择一种难度系数的动作, 当我们需要同时考虑运动员自己(personal)、回合(round)、裁判员(judger)、难度系数(difficulty)四者对得分的影响, 按照 d:(p × r) × j 模型进行数据分析时, 其方差分量无法估计, 这样难度系数成为了隐藏侧面(Brennan, R. L. 2001) , 故改用 p × r × j 进行 G 研究。为了评估难度系数在运动员之间、不同回合之间是否有显著差异, 我们也进行以难度系数为因变量, 运动员和回合为自变量的方差分析, 作为判断难度系数是否有必要进入考察范围的一个依据。在进行概化理论分析之前, 不需要对原始数据做任何特殊处理。D 研究将运动员的能力值作为测量目标, 进行 p×R×J 研究设计。难度系数作为一个隐藏侧面, 与运动员和回合两者的交互效应存在混淆(Brennan, 2001a)。为了让难度系数纳入视野, 所以同时也进行(p, R)×J 研究设计。这种设计优点在于, 难度系数纳入测量目标有一定合理性, 因为难度系数在一定范围内运动员是可以自选的, 难度系数本身也是运动员自身选择的结果。缺点在于将难度系数纳入测量目标的同时, 回合也纳入了测量目

* 1~4 组动作的号码均采用 3 位数。第一个数代表动作组别: 第二个数代表飞身动作(如果第二位数是“0”, 则表示没有飞身动作); 第三个数代表翻腾周数(以“1”为半周, “2”为一周, “3”为一周半, 以此类推)。第 5 组转体动作采用 4 位数。第一位数表示第 5 组(特指转体跳水); 第二位数表示翻腾转体的方向; 第三位数表示翻腾周数: 第四位数表示转体用数, 计算方法同前。第 6 组臂立动作也采用 3 位数。第一位数表示第 6 组 (特指臂立跳水); 第二位数表示臂立跳水的方向: 第三位数表示翻腾周数(计算方法同上)。例如“614”动作中“6”表示第 6 组臂立跳水, “1”表示采用第一组向前跳水方向翻腾, “4”表示翻腾两周。再如“632”, 是指第 6 组的臂立跳水动作, 用反身跳水方向翻腾一周。每个动作的号码以一个字母结尾。字母用来标示跳水空中姿势, 可以分为 A(直体)、B(屈体)、C(抱膝)、D(翻腾兼转体)4 种。例如: “107B”表示向前翻腾三周半屈体。“5253D”表示向后翻腾两周半转体一周半屈体。

标。尽管如此, 我们也可以通过这个设计将裁判员的影响与非裁判员的影响做一个二分。通过改变回合数和裁判数, 希望能找到比较好的测量设计。

采用项目反应理论的 MFRM 模型进行分析数据, 使用 John M. Linacre 编写的 Facets for Windows 3.54.1 版软件对数据进行数据处理。以各裁判员给各位运动员在各回合当中的得分为因变量, 同时考虑运动员自己、回合、裁判员、难度系数四个侧面及其二次交互效应对得分的影响。本次参赛选手共选择了难度系数为 3、3.2、3.3、3.4、3.5、3.6、3.8 共七个等级的难度系数, 其得分的观察值的最小值为 3.5, 最大值为 10, 等级间距为 0.5, 各等级上都有一定观测值分布。此外, MFRM 的被试能力参数的估计是独立于特定的题项的, 题项的难度系数的估计是独立于特定的被试样本的, 这就对测验的设计放松了要求, 即便是非平衡数据, MFRM 也能处理: 虽然每个运动员并没有选择所有的难度系数的动作, 但是由于裁判员对各难度系数下的运动员的表现都进行了评判, 所以裁判员的评分标准和运动员的表现将所有难度系数锚定在同一个参照框架下进行比较(O'Neill, Thomas, Lunz, 1996)。

3　结果

3.1　概化分析

3.1.1　G 研究　如表 1 所示, 总共有 7 个变异来源的方差分量得到了估计。根据这些方差分量的大小, 我们可以知道哪个来源的变异或误差最大, 并以此对测评进行改进。本研究中没有残差分量, 它们已经包含到三重交互效应当中去了。表 1 显示, 最大的变异来源为被测试者与回合的交互效应, 占总变异的 66.3%, 说明运动员在不同的回合中的得分的排序是不同的, 这一结果表明, 基于某一回合的数据对运动员的相对成绩的概化, 是不可靠的, 要准确估计被测者的成绩, 需要进行多个回合的比赛。其他来源从大到小依次为测量目标的主效应(20.96%), 被测者、回合与裁判三者的交互效应(8.32%), 回合(2.86%), 被测者与裁判的交互效应(1.36%), 回合与裁判的交互(0.48%), 裁判的主效应(0.00%)。

表 1　评分中各种变异来源的变异情况

变异来源	df	SS	MS	误差	方差分量	变异百分比
p	11	239.50	21.77	7.69	0.34	20.96
r	5	57.44	11.49	7.65	0.05	2.86
j	6	1.58	0.26	0.36	0.00	0.00
pr	55	415.70	7.56	0.13	1.06	66.30
pj	66	17.38	0.26	0.13	0.02	1.36
rj	30	6.76	0.23	0.13	0.01	0.48
prj	330	43.93	0.13	0.00	0.13	8.32

3.1.2　D 研究　奥运比赛最引人关注的是成绩排名, 因此我们更关心相对概化系数而非绝对概化系数。表 2 显示了在改变比赛回合数、裁判员数量的情况下, 相对概化系数的变动情况。从表 2 可以看到, 在 p×R×J 研究设计中, 当裁判数量为 7, 比赛回合数为 6 的情况下, 相对概化系数为 0.65, 不是很高。可以看到, 裁判员的数量的变化对估计的信度影响不大, 而回合的数量的变化对估计的信度影响比较大。与此同时, 我们应该注意到, p×R×J 设计中存在一个隐藏侧面, 即难度系数。同一个运动员在多个回合里的所完成的任务的难度系数不一定相同, 不同运动员在同一个回合里所选择完成的任务的难度也不定相同。方差分析也表明, 不同运动员之间的难度系数有显著差异, $F(11,487)=11.27$, $p<0.001$, 在不同回合之间的难度系数也有显著差异, $F(5,487)=115.55$, $p<0.001$)。测量目标和回合两者的交互效应与难度系数之间存在着混淆时, 使得 pr 的方差分量还包括难度系数引起的变异, 这样导致计算相对概化系数时的分母增大, 所以, 在 p×R×J 设计中, 相对概化系数存在被低估的情况。由于无法估出难度系数引起的变异的大小, 难以将隐藏侧面的作用纳入考虑范围。本研究也反映了概化理论的一个缺点, 即并不是总能够考虑所有侧面的作用。表 2 的下半部分的(p, R)×J 研究设计中, 可以看到, 裁判员的个数同样地对 g 系数影响不大, 在仅将裁判员因素作为误差时, g 系数为 0.99, 非常之高。考虑到难度系数的影响时, 实际的 g 系数应该没有这么高。

表 2 D 研究

		裁判员个数									
		1	2	3	4	5	6	7	8	9	10
回合数		g 系数									
1	pRJ 设计	0.23	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
2		0.37	0.38	0.38	0.38	0.38	0.38	0.39	0.39	0.39	0.39
3		0.46	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
4		0.53	0.54	0.55	0.55	0.55	0.55	0.55	0.55	0.56	0.56
5		0.58	0.59	0.60	0.60	0.61	0.61	0.61	0.61	0.61	0.61
6		0.62	0.63	0.64	0.64	0.65	0.65	0.65	0.65	0.65	0.65
7		0.65	0.67	0.67	0.68	0.68	0.68	0.68	0.68	0.68	0.68
8		0.67	0.69	0.70	0.70	0.71	0.71	0.71	0.71	0.71	0.71
9		0.69	0.71	0.72	0.73	0.73	0.73	0.73	0.73	0.73	0.73
10		0.71	0.73	0.74	0.75	0.75	0.75	0.75	0.75	0.75	0.75
11		0.72	0.75	0.76	0.76	0.77	0.77	0.77	0.77	0.77	0.77
12		0.73	0.76	0.77	0.78	0.78	0.78	0.78	0.78	0.78	0.79
	(p,R)J 设计										
1		0.90	0.95	0.96	0.97	0.98	0.98	0.98	0.99	0.99	0.99
2		0.91	0.95	0.97	0.97	0.98	0.98	0.99	0.99	0.99	0.99
3		0.91	0.95	0.97	0.98	0.98	0.98	0.99	0.99	0.99	0.99
4		0.91	0.96	0.97	0.98	0.98	0.98	0.99	0.99	0.99	0.99
5		0.92	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
6		0.92	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
7		0.92	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
8		0.92	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
9		0.92	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
10		0.93	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
11		0.93	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
12		0.93	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99

3.2 多面 Rasch 分析

3.2.1 运动员 表 3 输出的是 12 位参赛运动员的能力值的估计, Rasch 模型输出的能力值是等距量尺(Embretson, et al, 2000), 单位为 logit。表 3 中, 10 号运动员和 11 号运动员的 infit 和 outfit 都大于 1.5, 小于 2.0, 说明这两个运动员的得分的变异大于模型所期望的程度, 可能存在发挥失常或超常或其他情况, 但其得分的变动情况并没有太离谱, 属于意料之外, 情理之中。运动员侧面的分离信度为 0.98, 分离比为 6.79, 对运动员的成绩的差异程度进行卡方检验, $\chi^2_{(11)} = 517.5$, $p<0.001$, 表明运动员之间的得分差异显著, 这种差异是稳定、可重复的。

实际比赛中, 运动员最终的名次排序是先根据各回合的得分剔除两个最高分、两个最高分之后加总, 然后与难度系数相乘得到回合分, 并根据六个回合的总分进行名次排序。而 MFRM 使用的分数是没有经过剔除、也没有经过加权的原始分数, 当然, 在实际计算过程中, 综合考虑了裁判、难度等其他各个侧面。比较有意思的是, 表 3 的排名与 2008 奥运会决赛的实际排名(由高到低为: 11、10、9、2、12、5、7、4、3、8、1、6)相关非常显著(r=0.979, p<0.001), 但也有六名运动员的排序有差异, 这种差异也出现在一些重要名次上。

3.2.2 裁判员 表 4 输出的是裁判员的 MFRM 分析结果。就裁判员打分的宽严来看, 从第二列数据我们可以看出, 2 号裁判员打分最严苛, 3 号裁判最宽松, 两者相差 0.1 logits。总体上裁判的宽严差别都不是很大, 宽严度的标准差(0.03 logits)就相当地小。就裁判员打分的保守程度来看, 从 infit 和 outfit 的大小可以看出, 6 号裁判员打分时, 比较保守, 给出的分数变动幅度最小, 1 号裁判的 infit 和 outfit 都大于 1.2, 说明其打分比较冒进, 给出的分数变动幅度较大, 但总的来说, 裁判打分的变动幅度还是相当地合理和稳定。

表 3　运动员的 MFRM 分析

ID	运动员	能力值	*SE*	Infit MS	Outfit MS
11	Mathew Mitcham	1.29	0.10	1.65	1.62
9	Gleb Galperin	1.05	0.09	1.37	1.40
10	Zhou Luxin	1.01	0.09	1.70	1.62
2	Jose Antonio Guerra Oliva	0.56	0.08	1.35	1.36
12	Huo Liang	0.51	0.08	0.67	0.65
7	Thomas Daley	0.19	0.07	0.56	0.51
5	Mathew Helm	0.19	0.07	0.58	0.49
4	Rommel Pacheco	0.10	0.07	0.71	0.68
3	Patrick Hausding	0.00	0.05	1.32	1.25
1	Juan Guillermo Uran	–0.13	0.05	0.64	0.74
8	David Boudia	–0.16	0.05	0.73	0.71
6	Thomas Finchum	–0.24	0.05	1.08	0.90
Mean		0.36	0.07	1.03	0.99
SD		0.50	0.02	0.41	0.41

注：RMSE = 0.07, Adj *SD* = 0.49,　分离比 = 6.79, 分离信度 = 0.98, $\chi^2_{(11)} = 517.5$, $p<0.001$

表 4　裁判员的 MFRM 分析

裁判员	宽严度	*SE*	Infit MS	Outfit MS
2	0.04	0.05	0.81	0.87
5	0.02	0.05	0.91	1.14
4	0.02	0.05	0.93	0.98
7	0.02	0.05	1.00	1.06
1	0.01	0.05	1.23	1.28
6	-0.04	0.05	0.77	0.74
3	-0.06	0.05	0.90	0.90
Mean	0.00	0.05	0.93	1.00
SD	0.03	0.00	0.14	0.17

注：RMSE = 0.05, Adj *SD* = 0.00,分离比< 0.01, 分离信度<0.01, $\chi^2_{(6)} = 3.3$, $p = 0.77$

以上从裁判员个体打分的情况得出的结论, 在裁判员侧面的总体参数上也得到了印证。分离信度和分离比都小于 0.01, $\chi^2_{(6)} = 3.3$, $p= 0.77$, 说明没有因裁判的不同而导致运动员的得分差异显著。

3.2.3　难度系数　表 5 列出 MFRM 对不同难度系数的估计, MFRM 输出的难度值与原始数据的难度系数之间, 在排序上有些不同。难度等级为 3.5 和 3.3 的中等难度动作在 MFRM 估计出的难度排序中位置下移。此外, 第 4 个难度等级(3.4)的 outfit 为 1.27, 说明在这个难度上, 有异常分数出现。分离信度和分离比分别为 2.68 和 0.88, $\chi^2_{(6)} = 51.4$, $p< 0.001$, 表明在不同的难度系数上运动员的得分差异显著。

表 5　难度的 MFRM 分析

难度	测度值	*SE*	Infit MS	Outfit MS
3.8	0.23	0.05	0.94	1.19
3.6	0.19	0.05	0.77	0.69
3.4	0.07	0.03	1.19	1.27
3.2	0.04	0.04	0.93	1.09
3.5	–0.09	0.08	0.38	0.42
3.0	–0.19	0.07	0.72	0.80
3.3	–0.24	0.08	0.67	0.75
Mean	0.00	0.06	0.80	0.89
SD	0.17	0.02	0.23	0.28

注：RMSE = 0.06, Adj *SD* = 0.16, 分离比 = 2.68, 分离信度= 0.88, $\chi^2_{(6)} = 51.4$, $p < 0.001$

3.2.4 回合 回合, 实际上包含了时间因素。从表 6 的第二列可以看到, 比赛的趋势是越到后面进行得越艰难。结合 infit 和 outfit 的情况来看, 第一轮和第二轮各运动员发挥都还正常, 中间两轮(3 和 4)表现有点沉闷, 运动员之间的得分没有能够拉开应有的距离(infit 在 0.6 左右), 最后两轮比较激烈, 尤其是在第六轮比赛中, 运动员的得分差异最显著, 出现了超常发挥或意外失手的情况(outfit>1.2)。总的来看, 根据 MFRM 的估计, 回合的分离比为 2.96, 分离信度为 0.90, $\chi^2_{(5)}$ =50.3, p<0.001, 即不同回合之间的得分差异显著。

3.2.5 评定等级量表 无论是经典测量理论还是概化理论, 对评定等级量表的假设有两点: 评定量表的等级之间是等距的; 被测者特质水平越高, 得到的等级也越高。但这两点假设, 尤其是前者, 经常没有得到满足。从表 7 的第 5 列中我们可以看到, 原始数据中的等级间距并不是恒等的, 最大值为 0.34, 最小值为–0.08, 平均数为 0.13, 标准差为 0.11。其中等级 6.5 出现了等级倒置的情况, 说明裁判员没有很好地掌握这两个等级之间的差别, 没有能够正确地使用相应的等级来评定运动员的表现, 导致出现高分低能、低分高能的谬误。第 6、7、7.5 这三个等级的 Outfit MS 都大于 1.2, 说明在实际使用中, 这几个等级把握得不是很准确, 打分不够稳定。这对于诊断裁判员的打分模式及对他们进行培训, 都是极好的信息。

表 6 回合的 MFRM 分析

回合	测度值	*SE*	Infit MS	Outfit MS
6	0.17	0.04	1.26	1.55
5	0.13	0.04	1.09	1.17
1	0.02	0.05	0.81	0.76
4	0.00	0.04	0.60	0.61
3	–0.04	0.05	0.62	0.78
2	–0.28	0.06	1.00	1.05
Mean	0.00	0.05	0.90	0.99
SD	0.15	0.01	0.24	0.31

注：RMSE = 0.05, Adj *SD* = 0.14, 分离比 = 2.96, 分离信度 = 0.90, $\chi^2_{(5)}$ = 50.3, p < 0.001

表 7 MFFM 评定等级量表

等级	案例数	平均能力值	Outfit MS	等级间距	*SE*
3.5	3	–0.58	0.40	/	
4.0	7	–0.46	0.60	0.12	0.57
4.5	9	–0.34	0.80	0.12	0.33
5.0	8	–0.34	0.50	0.00	0.25
5.5	4	–0.25	0.70	0.09	0.22
6.0	16	–0.10	1.40	0.15	0.21
6.5	19	–0.18*	0.80	–0.08	0.18
7.0	37	0.03	1.30	0.21	0.15
7.5	55	0.23	1.40	0.20	0.13
8.0	115	0.26	1.20	0.03	0.11
8.5	107	0.38	1.20	0.12	0.11
9.0	65	0.72	0.90	0.34	0.12
9.5	47	0.98	0.90	0.26	0.15
10.0	12	1.16	0.90	0.18	0.29

注：*步校准错乱

3.2.6 交互效应 运动员与回合的交互效应有 $\chi^2_{(72)}$ = 643.1, p<0.001, 偏差案例占案例数的 36/72, 说明各运动员在不同回合发挥的程度变化显著。如图 1 所示*, 2 号运动员发挥一路走低, 10 号运动员前 5 个回合发挥平稳, 但最后一个回合跌入谷底, 而 9 号和 11 号运动员中间虽有小挫, 但基本上是越

*图 1 中的 Z 分数对软件输出的数字做了符号转换。

战越勇。难度与回合的交互效应有 $\chi^2_{(23)}$ =119.4, p< 0.001, 偏差案例占案例数的 13/23。运动员与难度的交互效应有 $\chi^2_{(57)}$ =476.7, p<0.001, 偏差案例数占总案例数的 32/57, 说明各运动员对难度系数不同的动作掌握的程度是有显著差异的。而裁判员与回合的交互效应不显著($\chi^2_{(42)}$ =13.2, p>0.999), 无一案例偏差显著; 裁判员与运动员的交互效应方面,偏差显著的案例占总案例数的 1/84, 即第三位裁判对 12 号运动员的打分偏高(z=−2.04), 但裁判员与运动员的交互效应总体不显著($\chi^2_{(84)}$ =44.6, p> 0.999)。

可以显见的是, MFRM 模型比概化理论能提供更多的信息。在交互效应分析方面, 我们不仅可以得到交互效应显著与否, 还可以看到交互作用的模式。比如, 5 号裁判给 10 号运动员的打分, 并没有显著的偏差(z=0.64), 但是打分的变动比较大(infit= 3.3, outfit=3.1), 这种情况的出现, 可能是裁判员打分过程中有一个自动纠错的过程。

分析报告没有提供四重交互作用的统计推断, 但提供了相关个案的偏差报告, 从表 8 我们可以看到几个偏离预期较严重的打分, 如 10 号运动员在第六回合进行难度系数为 3.4 的动作时, 5 号裁判和 1 号裁判给他的分数分别是 6 和 6.5, 而依这两位裁判的的宽严标准, 在正常情况下, 在第六回合的这个难度上, 两裁判都应给出 8.5 分。MFRM 提供的只是一个现象的描述, 至于原因, 需要结合实际情况才能给出合理的解释, 就两位裁判同时出现显著低判的情况而言, 该运动员在最后一个回合发挥失常的可能性比较大。

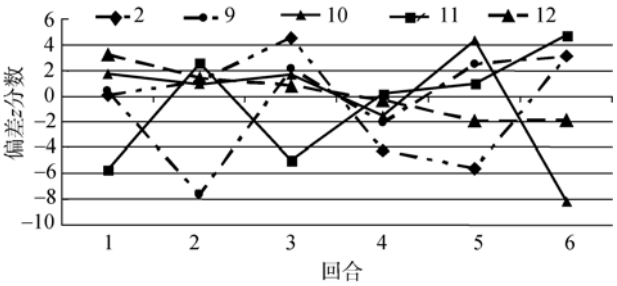

图 1　部分运动员各回合得分偏差分析

表 8　奇异值

实际得分等级	期望得分等级	残差	标准化残差	运动员	回合	难度系数	裁判员
7.0	9.0	−4.0	−3.0	9	2	3.2	7
6.0	8.5	−5.2	−3.0	10	6	3.4	5
6.5	8.5	−4.2	−3.0	10	6	3.4	1

4　讨论

本研究同时引入了 GT 和 MFRM, 为奥运会跳台跳水的运动员的得分情况进行了详尽的分析, 提供了丰富的信息。下面就运动员得分的总的变异来源情况, 测量目标和测量侧面的具体情况做进一步讨论。

4.1　得分的变异来源

本研究中, GT 分解出得分的 7 个变异来源, 其大小依次为: 运动员与回合的交互、运动员、运动员,回合及裁判三者的交互、回合、运动员与裁判员的交互、回合与裁判员的交互、裁判员。GT 并没有提供一个指标来判断侧面的效应是否显著, 而 MFRM 提供了分离比和分离系数, 并提供 p 值判断各侧面的效应是否显著。MFRM 认为运动员与回合的交互、运动员、回合效应显著, 而运动员与裁判员的交互、回合与裁判员的交互、裁判员效应不显著。另外, MFRM 还分离出难度系数的效应, 并认为其作用显著。而概化理论分析中, 难度系数作为隐藏侧面, 其效应没有像 MFRM 中那样得到估计, 说明概化理论现在虽然已经能处理一部分非平衡设计的数据, 但其对平衡数据的偏爱并没有完全消除(Brennan, 2001 b); 在概化分析中, 运动员、回合、裁判三者的交互效应的方差分量占到总变异的 8.32%, MFRM 本次没有估计这三者的三重交互效应, 但提供了存在偏差的个案信息。除涉及难度系数部分有差异外, 可以认为, 总的来说 GT 与 MFRM 的估计是比较一致的。实际上这两者的估计并不总是一致, 如 Lynch 等(1998)的研究中就出现的 GT 与 MFRM 估计不一致的现象。Linacre(1993)认为, MFRM 估计的信度一般会更可靠, 因为 MFRM 各侧面信度的估计是在测度已经做了调整的基础上进行的, 使用的是排除了其他侧面方差之后的自己的误差项, 而 GT 是基于原始数据计算, 同时还容易受到分数分布的影响。

4.2　运动员

用 GT 的术语来说, 运动员是测量目标, MFRM 则将其也看作测量侧面。GT 估计出的与运动员相关的变异占总变异的 96.94%(参见表 1), 其中, 运动员与回合的交互、运动员, 回合及裁判员三者的

交互, 都是运动员得分变异的重要来源, 所造成的变异百分比都大于 8%。与回合的交互表明, 在运动员体力和运营成本容许的情况下, 适当增加回合数可以得到更加稳定的运动员的能力的估计。

MFRM 诊断出运动员与回合的交互作用、运动员与难度系数的交互作用显著, 表明运动员在不同的回合得分波动比较大, 有的越战越勇, 有的却功败垂成。此外, 运动员在完成不同难度系数的动作时表现不一。这些信息对 GT 分析既是印证, 又是补充。这些信息对于有针对性的训练和运动员心理调适也有非常重要的意义。

在实际评分中, 通过剔除两个最高分和最低分的办法, 得出运动员的成绩排序, 这对舞弊起到了一定的遏制作用, 但也容易造成另一种形式的不公平, 比如得分“8、8、8、8、8、7、7”和“9、9、8、8、8、8、8”, 剔除两个高分和低分后, 两者得分相同, 但实际上, 前者的真实能力可能比后者高, 而这种差异被计分规则给抹杀掉了。实际评分中根据难度系数对分数加权措施有一定合理性, 但这些措施又过于简单, 难度系数与难度的实际测度有差异。此外, 实际评分并没有考虑到评定等级量表等因素的影响, 所依据的原始分数仅是等级数据。MFRM 依据的数据既没剔除两个最高分和最低分, 也没对难度系数进行加权, 这正是因为 MFRM 本身具备甄别裁判偏误和难度系数的影响的能力, 并提供剔除了难度系数、回合及裁判偏差等具体测量情景影响之后每位运动员的能力估计值。MFRM 得出的运动员成绩排名与奥运实际名次显著相关, 但也有一半运动员在两套办法中得出的名次是不同的。从以上分析可以看到, MFRM 的估计应该是更加准确可信。

4.3 其他侧面

裁判员评分是否公允, 是所有关心体育比赛的人所关注的一个焦点。本研究中, GT 和 MFRM 都同时认为裁判员主效应或裁判与运动员的交互效应都很小, 也就是说, 裁判员的打分是公允的, GT 的 D 研究表明增减裁判员数对测评信度影响不大。MFRM 对评定等级量表的分析发现, 裁判员对得分区间的部分分数段存在盲点, 对几个等级把握不准, 甚至出现等级倒置的情况, 仍有必要对之进行有针对性的培训。

难度系数越高, 动作越难掌握, 但得分权重却越高。GT 未能分离出难度效应, MFRM 分析表明, 难度效应显著, 有意思的是, MFRM 对难度进行赋值时还发现, 难度系数为 3.5 和 3.3 的动作, 实际难度有大幅下降, 甚至分别在难度系数为 3.2 和 3.0 的动作之下。应该说, 这是有实际意义的, 从得分的角度来说, 运动员加强中等难度及中上等难度的动作的练习, 是一个很好的得分策略。这种实际难度系数的变化是否是因为中等难度的动作在比赛时诱发了中等程度的紧张而导致发挥普遍地好, 还是因为这些动作平时得到了更多的重视而致使实际难度下降, 尚待进一步探究。

在 GT 分析中, 回合的主效应解释了变异的 2.86%, 而与运动员的交互效应解释了变异的 66.30%, 所以, 从 GT 的角度来说, 增加比赛的回合数, 是提高信度的最佳策略。MFRM 更是能提供所有运动员在各个回合偏差模式, 对于分析运动员心理并在训练中对之进行辅导提供了有益信息。

4.4 GT 与 MFRM 之比较

通过前述结果及讨论, 我们可以了解到, 概化理论和项目反应理论在计算各个侧面的变异的相对大小方面是比较一致的, 但这两者测评技术处理变异来源的方式等方面却大相径庭。首先, 概化理论分析数据时, 假定输入的数据资料就是等距数据, 而实际上可能仅是等级数据, 输出的运动员的得分的平均数、全域分数与输入数据的尺度是一样的; 项目反应理论的一个优势就是能给出一个等距的测定。其次, 概化理论能估计出的效应都是侧面层的, 并由此推论在类似情况下测量的可信度, 给出的测量改进建议更多地体现在宏观上的意见, 关注测量侧面的水平数的变化及测量模式对测量信度的影响。项目反应理论不仅提供侧面层的效应显著与否, MFRM 的关注点更主要地是在每个侧面的每个元素(被测者、裁判等)上, 它能给出各个侧面上所有元素的估计值及误差。MFRM 在调整量尺、给出个体层的估计值、诊断裁判员打分的模式等方面比较有优势。这种方法在人员选拔(包括考官的选拔)中非常实用, 就本研究而言, MFRM 所揭示的难度系数排序偏移、回合效应、评定等级量表的错乱等现象都值得运动心理学家及相关部门进一步研究。

5 结论

本研究使用 GT 和 IRT 中的 MFRM 对 2008 年北京奥运会男子十米跳台跳水决赛数据进行分析, 发现比赛评分尚有可以改进的余地和空间, 为运动员、裁判员的选拔与培训提供了有用的信息。同时,

通过两种方法的运用, 我们也可看到他们的共同之处及各自的优缺点。

(1) 运动员的成绩不仅仅受运动员自身的影响, 同时也受到回合与难度系数等测量情景因素的影响。GT 分解出得分的变异来源, 但并没有提供一个指标来判断侧面的效应是否显著; MFRM 提供了分离比和分离系数, 并提供 p 值判断各侧面的效应是否显著。总的来说, 本研究中, GT 与 MFRM 对各侧面效应大小的估计是比较一致的。但 GT 的目的是通过分解变异来源, 计算概化系数, 而 MFRM 是要剔除其他侧面的影响, 对相应侧面的要素做出估计。

(2) MFRM 发现动作难度是影响运动员成绩的一个重要方面; 同时, 不同难度系数的动作的实际难度有所变化。为运动员平常训练及心理辅导提供了有益的信息。而 GT 无法分离难度系数的效应。MFRM 的锚定功能使其比 GT 能更加灵活地适应各种数据。

(3) MFRM 分析认为各裁判员打分的宽严度差异不显著, 裁判员与运动员的交互作用不显著, 即不存在不公正的现象, 但也存在个别裁判给个别运动员打分时宽严度波动较大的情况; GT 发现裁判数量的多少对运动员得分影响非常有限, 通过增加裁判数来提高概化系数, 意义不大。MFRM 提供了丰富的要素层面的信息, 能为裁判的选拔和培训提供有针对性的信息。GT 则侧重宏观层面的信息, 为测量设计提供理论依据。

(4) MFRM 和 GT 同时认为运动员与回合的交互作用对成绩有重要影响。MFRM 勾勒出各运动员与回合交互的不同模式及不同回合上运动员得分差异的变化模式, 不仅说明了运动员状态的调试的重要性, 也为有针对性地提供运动员心理辅导提供了依据。GT 分析发现, 增加比赛回合的次数, 是提高测量信度的好办法。

(5) MFRM 表明本次比赛评分中存在步校准错乱现象, 这对于识别、纠正裁判员打分过程中对标准把握不准的情况, 有重要的现实意义。MFRM 对评定等级量表重新进行锚定, 将得分锚定在连续量表上, 而 GT 处理的, 只是假定为连续数据, 而实际上只是等级数据, 各等级之间的间距并不恒等。

(6) GT 分析实际的测量情景关系, 通过不同的确定侧面水平数的做法, 有针对性地揭示不同情况下概化系数, 侧重外部效度问题; MFRM 则着重被试的能力水平、裁判员的偏差等的估计, 给各侧面的要素以更精准的估计, 侧重内部效度问题。两者侧重的方面有所差异, 分别在不同的侧面为测评工作提供诊断资讯。实际应用中, 可以根据实际需要联合运用或择取其一。

值得注意的是, 对于所有的裁判员集体作弊, 故意给某个被测者打高分或打低分的情况, 包括 GT 和 MFRM 在内的所有统计手段, 都会束手无策, 所以, 主观评分的改进, 不仅需要借助于统计手段, 但又不能完全依赖于统计手段。

致谢: 感谢江西师范大学心理学院戴海崎教授和两位匿名评审专家给本文所提的建设性意见。

参 考 文 献

Brennan, R. L. (2001a). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L. (2001b). *Manual for urGENOVA*. Iowa City, IA: Iowa Testing Programs, University of Iowa.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The *dependability of behavioral measurements. theory of generalizability for scores and profiles*. New York: Wiley.

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: Erlbaum.

Engelhard, G. (1994a). The measurement of writing ability with a many - faceted Rasch model. Applied Measurement in Education, 5,171–191

Engelhard, G. (1994b). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112

Everett V. Smith, JR., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological* Measurement, *64*(4), 617

Linacre, J. M. (1993). *Generalizability Theory and Many-facet Rasch Measurement*. Paper presented at the Annual Meeting of the American Educational Research Association.

Linacre, J. M., Engelhard G., Jr., Tatum, D.S., & Myford, C.M. (1994). Measurement with judges: Many- faceted conjoint measurement. *International Journal of Educational Research*, *21* (6): 569–577

Linacre, J. M. (2003). A user's guide to FACETS [computer program manual]. Chicago: MESA Press.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180.

O'Neill, Thomas R., & Lunz, Mary E. (1996). *Examining the invariance of rater and project calibrations using a multi-facet rasch model.* Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996)

Qi, S. Q., Dai, H. Q., & Ding, S. L. (2002). *Modern Education and Psychometric Principles* (in Chinese). Beijing: High Education Press.

[漆树青, 戴海崎, 丁树良. (2002). *现代教育与心理测量学原理*. 北京: 高等教育出版社.]

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*, 239–261.

Sun, X. M., & Zhang, H. C. (2006). An IRT analysis of rater bias in structured interview of national civilian candidates (in Chinese). *Acta Psychologica Sinica, 38* (4), 614–615

[孙晓敏, 张厚粲. (2006). 国家公务员结构化面试中评委偏差的 IRT 分析. *心理学报*, *38* (4), 614–615.]

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*(1), 35–51.

Yang, Z. M., & Zhang, L. (2003). *Generalizability Theory and Its Applications* (in Chinese). Beijing: Educational Science Publishing House.

[杨志明, 张雷 (2003). *测评的概化理论及其应用*. 北京: 教育科学出版社.]

A Comparison of GT and IRT: An Analysis of Performance Rating of Men's 10 Meters Platform Diving in Beijing Olympic Games

YU Zong-Huo[1], TANG Xiao-Juan[2], WANG Deng-Feng[1]

([1] *Department of Psychology*, *Peking University*, *Beijing* 100871, *China*)

([2] *College of Mathematics and Information Science*, *Nanchang Hangkong University*, *Nanchang* 330063, *China*)

Abstract

Generalizability Theory (GT) and Item Response Theory (IRT) have improved the Classical Test Theory (CTT) in different aspects. They put focus on macro-level and micro-level of measurement, respectively. Both GT and Multi-Facet Rasch Measurement model (MFRM, which is one case of IRT methods) can be applied to decompose the variances from different sources (including error) in the Performance Rating and to estimate the reliability of rating. The results from both of them can give researchers some recommendations about how to improve the Performance Rating. This paper tries to find how they perform differently in the way of improving the rating process in Beijing Olympic Games through making a comparison between GT and MFRM.

Those athletes' scores from 10 meters platform diving in Beijing Olympic Games form the data to be analysis. In the 2008 Beijing Olympic Games, there were twelve athletes who participated in the final of Men's 10 meters platform diving. Each athlete dived six times, and was marked independently by seven referees each time. In total, there are 12×6×7=504 data points. Based on this dataset, both GT and MFRM are applied to analyze four facets (including round, person, referee, and difficulty) of these scores. However, as a hidden facet, difficulty can't be separated in GT.

The results from GT and MFRM suggest consistently that the athlete, the round, and their interaction are important sources of variation in these scores, and that the referees have not significant contribution to variance in athletes' scores. At the same time, the results from MFRM indicate that the difficulty is also a significant source of variation. Based on these results, we can find some ways to improve scoring from different aspects. For example, we find that the g coefficient is influenced significantly not by the number of referee but by the number of rounds. Therefore, it's helpful to improve the reliability of rating through increasing the number of rounds. MFRM gives the measure of individual elements within each facet, the standard errors for each element and the diagnostic fit statistics to detect aberrant responses. Based on the analysis of MFRM, We find the referees disordered the step calibrations of the scale around the category of 6.5. The results from MFRM also give birth to a new ranking which is really different from that given in the 2008 Beijing Olympic Games.

In sum, we find that GT and MFRM are consistent totally in estimating the sources of variation. However, both methods have their own advantages. GT is more helpful in the way of design of measurement, and MFRM is more helpful in the ways of measure of individual elements within each facet and detecting aberrant responses. Moreover, MFRM can separate the effects of round, referee, and difficulty more successfully and produce a more precise estimation of ranking of athletes than the method used in 2008 Beijing Olympic Games.

Key words generalizability theory; multi-facet Rasch measurement; performance rating