文章编号:1000-5862(2019)04-0368-08

多维标准参照测验下分数报告质量评价指标

宋丽红[1],汪文义[2]
(1. 江西师范大学初等教育学院,江西 南昌 330022;2. 江西师范大学计算机学院,江西 南昌 330022)

摘要: 标准参照测验主要关注学生在特定内容、知识或技能上的掌握程度和表现水平. 分数报告中表现水平的分类信度和效度,通常采用分类一致性和分类准确性进行评价. 首先介绍多维测验下的分类决策规则;然后介绍多维项目反应理论模型下 3 类分类一致性和分类准确性指标,一类是基于总分量尺的指标、另外 2 类分别是基于似然函数和信息矩阵定义在能力量尺的指标;同时还介绍了这些指标的作用;最后指出分类一致性和分类准确性可以用于评价标准参照测验子分数的分类信度和效度,还可以指导计算机分类测验选题和组卷.

关键词: 多维项目反应理论;分数报告;决策规则;分类准确性;分类一致性
中图分类号: B 841.7 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2019.04.07

0 引言

标准参照测验(criterion-referenced tests,CRT)主要报告学生在特定内容、知识或技能上的掌握程度和表现水平. 标准参照测验一般在各个维度上将被试分为 2 个水平(掌握、未掌握)或 3 个水平(初级水平、熟练水平、高级水平)等表现水平. 根据 CRT 分数报告结果及结果解释,教师可改进教学侧重点,学生也可根据自己的强项和弱项进行针对性学习. 因此,CRT 有助于发挥考试的诊断功能和促进学生个性化学习[1]. 标准参照测验已经广泛应用于水平、资格和成就考试等,例如国际学生评价项目、国际阅读素养测评项目、国际数学和科学成就趋势调查、美国国家教育进步评价、美国研究生入学考试、中国国家基础教育质量监测等[2-3]. 随着新课程标准的建立,基于新课程标准的标准参照测验也有待开发. 因为一旦建立和采用新标准,就需要开发新测试,以测量学生是否达到相关标准[4].

任何测量都存在测量误差. 测验信度和效度指标可用于评价测量各种随机和系统误差大小. CRT 通常会估计和报告学生在特定内容、知识或技能上的子分数或能力分数,再结合专家划定的标准或划界分数,给出学生的表现水平. 一般而言,测验题量、难度分布、题目质量、测量模型、子分数或能力分数估计方法等均会影响表现水平的分类信度和效度,并且分类结果会影响分数报告使用者的决策. 因此,表现水平分类结果的稳定性和准确性对于分数报告十分重要. 分类一致性和分类准确性指标,成为研究者关注的重点[5-6]. 作为信度指标的分类一致性,它是指 2 次重复测量中被试观察分类或表现水平一致的比率,主要衡量分类结果的稳定性. 作为效度指标的分类准确性是指被试观察分类与其潜在真实分类相同的比率[6-7].

下面先简要介绍分类一致性和分类准确性指标的发展概况[8-15]. 最早是采用平行测验的方式来估计分类一致性和分类准确性. 因为平行测验在实际中较难实现,后来有研究考虑如何从单个测验数据估计分类一致性和分类准确性. 随着单维和多维项目反应理论(multidimensional item response theory,MIRT)的发展,基于经典测验理论分类一致性和分类准确性指标,逐渐推广并应用于项目反应理论下指标估计. 考虑到项目反应理论的优势,本文主要关注单个测验和项目反应理论模型下分类一致性和准确性指标及其估计方法. 按照分数报告所采用的量尺不同,这些指标主要分为 2 类[9]:基于观察分数(测验总分)的决策指标和基于潜在能力分数的决策指标. 其中,基于观察分数的决策指标主要采用 W. C. Lee[7,11] 提出的方法进行估计,基于潜在能力分数的决策指标主要采用 Guo Fanmin[8] 或 L. M.

收稿日期:2019-02-17
基金项目:江西省教育科学“十二五”规划一般课题(13YB032)资助项目.
作者简介:宋丽红(1981-),女,江西新干人,副教授,博士,主要从事教育测量研究. E-mail: viviansong1981@163.com

Rudner[13] 提出的方法进行估计.

文献[7-8]的方法开始主要用于单维项目反应理论模型下指标估计. 众多实证研究发现,前面提到的许多大型标准参数测验均为多维测验[16-20]. 这极大地推动了 MIRT 相关理论和应用研究迅速发展[21-28]. 伴随着 MIRT 的发展,对于多维测验,有些研究[14-15]采用文献[7]的方法估计不同内容维度分数的分类一致性和分类准确性,其采用的 MIRT 模型主要有简单结构多维模型、双因子模型和题组模型. 近年来有些研究[3]发现,基于能力分数指标比基于观察分数指标更高. 因此,最近一些研究者[29-31]将基于能力分数的文献[8]的方法、文献[13]的方法推广到 MIRT 模型,并比较了各方法的表现. 本文在介绍多维项目反应理论模型之后,重点介绍分类决策规则以及 3 类分类一致性和分类准确性指标.

1　多维等级反应模型

下面先简要介绍后面要使用的多维等级反应模型(multidimensional graded response model,MGRM). MGRM 是等级反应模型的多维模型,是多维能力下有序多值评分项目的测量模型. 约定以下记号: 样本中被试数为 N,即被试 $i=1,2,\cdots,N$; 测验项目数为 J,即项目 $j=1,2,\cdots J$; 项目 j 的最低分数等级为 0,最高分数等级为 K_j,对应等级分数 $k=0,1,\cdots,K_j$; 被试 i 在项目 j 的得分记为 y_{ij},它的取值为 $0\sim K_j$ 的整数; 测验结构的潜在能力维度记为 d; 被试 i 的潜在能力列向量记为 $\boldsymbol{\theta}_i=(\theta_{i1},\theta_{i2},\cdots,\theta_{id})'$; $\boldsymbol{\alpha}_j$ 表示项目 j 与区分度有关的参数向量; β_{jk} 是项目 j 的第 k 个等级难度,它满足严格递增关系 $\beta_{j1}<\beta_{j2}<\cdots<\beta_{jK_j}$. 若采用双参数 Logstic 模型,则能力为 $\boldsymbol{\theta}_i$ 的被试 i 得分为 k 及以上分数的概率为

$$P^*_{jk}(\boldsymbol{\theta}_i)=P(y_{ij}\geqslant k\,|\,\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)=1/(1+\exp(\beta_{jk}-\boldsymbol{\alpha}_j\boldsymbol{\theta}_i)),$$

其中 $j=1,2,\cdots,J,k=1,2,\cdots,K$. β_{jk} 越小表示被试越容易得到等级分数为 k 或更高等级分数. 该模型假设 $P(y_{ij}\geqslant 0\,|\,\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)=1$ 和 $P(y_{ij}\geqslant K_j+1\,|\,\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)=0$,且项目 j 的各个等级难度是严格单调递增. 由此可知,能力为 $\boldsymbol{\theta}_i$ 的被试 i 恰得 k 分的概率等于得 k 分或更高分的概率与得 $k+1$ 分或更高分的概率之差:

$$P_{jk}(\boldsymbol{\theta}_i)=P(y_{ij}=k\,|\,\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)=P^*_{jk}(\boldsymbol{\theta}_i)-P^*_{jk+1}(\boldsymbol{\theta}_i),\quad(1)$$

其中 $k=0,1,2,\cdots,K_j$.

MGRM 定义了给定能力为 $\boldsymbol{\theta}_i$ 的被试 i 在项目 j 上作答反应为 y_{ij} 的条件分布. 已知作答反应矩阵或得分阵,有计算机程序或软件(包)可用于多维模型的项目参数和被试能力估计[32],如 BMIRT(bayesian multivariate item response theory)、IRTPRO 软件和 R 软件下 mirt 包等. 在局部独立假设下,给定项目参数估计($\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$)与观察数据 y_i,最大化下面似然函数可得到被试能力估计:

$$L(y_i\,|\,\boldsymbol{\theta}_i,\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{j=1}^{J}\prod_{k=0}^{K_j}P(y_{ij}=k\,|\,\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)^{I_{(y_{ij}=k)}},\quad(2)$$

其中示性函数定义为

$$I_{(y_{ij}=k)}=\begin{cases}1, y_{ij}=k,\\ 0, \text{其他}.\end{cases}$$

2　决策规则

决策规则直接影响测验分类结果的信度和效度. 根据教育与心理测量标准,对于学生有重要影响(如升学、录取)的决策,不能仅基于单个方面的测验分数做决策[33],而要求使用多重测量(multiple measures)结果做决策,以提高测量的信度、效度、公平性等[34-35]. 多重测量结果一般按照一定决策规则生成合成分数(composite score). 合成方法通常可采用联合(conjunctive)、补偿(compensatory)、联合和补偿混合、验证(confirmatory)规则. 相关规则已经应用于英语考试、通识考试和学业评价等[33-34,36-37]. 其中,联合规则要求被试在各个测量目标上达标,补偿规则允许测量结果之间补偿,验证规则用于用一个测量去证实或评估其他测量结果的质量. 研究生入学考试同时规定考试科目单科分和总分最低要求,这属于混合型决策规则. MIRT 能细致地反馈学生在各个内容、知识和技能方面的信息,它特别适合于分析和合成多重测量结果[25,38]. 下面主要在 MIRT 框架下介绍 3 种多维潜在能力下的决策规则[29-31].

1) 基于各个能力分数的决策规则,决策区域定义如下:

$$R_{1k}=\{\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_d)\,|\,\tau_{0k}<\theta_k<\tau_{1k},-\infty<\theta_{k'}<+\infty,k'=2,3,\cdots,d\},$$

$$R_{hk}=\{\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_d)\,|\,\tau_{(h-1)k}\leqslant\theta_k<\tau_{hk},-\infty<\theta_{k'}<+\infty,k'=1,\cdots,k-1,k+1,\cdots,d\},$$

其中 $h=2,3,\cdots,H,\tau_{hk}$ 为第 k 维能力分数量尺上的划界分数,满足 $-\infty=\tau_{0k}<\tau_{1k}<\cdots<\tau_{Hk}=+\infty$.

2) 基于合成能力分数的决策规则,决策区域定义如下:

$$R_{1(H+1)} = \{\boldsymbol{\theta} = (\theta_1,\theta_2,\cdots,\theta_d) \mid \tau_{0(H+1)} < \sum_{k=1}^{d} w_k\theta_k < \tau_{1(H+1)}\},$$

$$R_{h(H+1)} = \{\boldsymbol{\theta} = (\theta_1,\theta_2,\cdots,\theta_d) \mid \tau_{(h-1)(H+1)} \leqslant \sum_{k=1}^{d} w_k\theta_k < \tau_{h(H+1)}\},$$

其中 $h = 2,3,\cdots,H$,w_k 表示第 k 维能力上的权重,$\tau_{(h-1)(H+1)}$ 表示合成能力分数量尺上的划界分数,满足 $-\infty = \tau_{0(H+1)} < \tau_{1(H+1)} < \cdots < \tau_{H(H+1)} = +\infty$.

3) 基于各个能力和合成分数的决策规则,决策区域定义如下:

$$R_h = \{\boldsymbol{\theta} = (\theta_1,\theta_2,\cdots,\theta_d) \mid \tau_{(h-1)k} \leqslant \theta_k, k = 1,2,\cdots,d, \tau_{(h-1)(H+1)} \leqslant \sum_{k=1}^{d} w_k\theta_k\} - \bigcup_{h'=h+1}^{H} R_{h'}.$$

3 分类一致性和分类准确性

3.1 基于文献[7]方法的分类一致性和分类准确性指标

记 $g(\boldsymbol{\theta})$ 表示能力分布的密度函数. 根据测验总分将被试分为 H 类(或表现水平),设置划界分数或划界点:$s_0,s_1,\cdots,s_H$,满足 $0 = s_0 < s_1 < \cdots < s_{H-1} < s_H = +\infty$ 且 $s_{H-1} < \sum_{j}^{J} K_j$. 当被试观察总分 $< s_1$ 时,被试判为第 1 类;当 $s_1 \leqslant$ 被试观察总分 $< s_2$ 时,被试判为第 2 类;依次类推,当被试观察总分 $\geqslant s_{H-1}$ 时,被试判为第 H 类.

3.1.1 **分类一致性指标** 被试的测验总分随机变量 X 的概率分布为

$$P_r(X = x) = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} P_J((X = x) \mid \boldsymbol{\theta}) \cdot g(\boldsymbol{\theta})\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_d,$$

其中随机变量 X 的观察值 $x = \sum_{j=1}^{J} y_j$,它表示被试在测验总分的可能取值,且 $0 \leqslant x \leqslant \sum_{j=1}^{J} K_j$,$P_J(X = x \mid \boldsymbol{\theta})$ 表示能力为 $\boldsymbol{\theta}$ 的被试在含 J 个项目的测验总分为 x 的条件概率. 在项目反应理论的局部独立假设成立情况下,对于测验长度为 J、能力为 $\boldsymbol{\theta}$ 的被试在测验上总分为 x 的条件概率的递推公式为

$$P_J(X = x \mid \boldsymbol{\theta}) = \sum_{k=0}^{\min(K_J,x)} P_{J-1}(X = x-k \mid \boldsymbol{\theta}) \cdot P_{Jk}(\boldsymbol{\theta}), \quad (3)$$

$P_{Jk}(\boldsymbol{\theta})$ 由(1)式计算,它表示能力为 $\boldsymbol{\theta}$ 的被试在项目 J 恰得 k 分的概率,$P_{J-1}(X = x-k \mid \boldsymbol{\theta})$ 表示前 $J-1$ 个项目上总分为 $x-k$ 的概率. (3)式也可以写成容易理解的公式:

$$P_J(X = x \mid \boldsymbol{\theta}) = \sum_{y_1,y_2,\cdots,y_J;\sum_{j=1}^{J} y_j = x;0 \leqslant y_j \leqslant K_j;j=1,2,\cdots,J} \prod_{j=1}^{J} P_{jy_j}(\boldsymbol{\theta}), \quad (4)$$

(4)式表示给定能力 $\boldsymbol{\theta}$ 下的所有满足测验总分为 x 的所有可能得分向量$(y_1,y_2,\cdots,y_J)$ 的联合概率或似然函数之和.

根据给定能力 $\boldsymbol{\theta}$ 下测验总分 X 的条件分布、决策规则中指定的观察分数量尺上的划界分数,可以得出能力为 $\boldsymbol{\theta}$ 的被试测验总分 X 位于表现水平第 h 类所在区间的概率,即能力为 $\boldsymbol{\theta}$ 的被试被分到第 h 类表现水平的概率为

$$p_{\boldsymbol{\theta}}(h) = P_J(s_{(h-1)} \leqslant X < s_h \mid \boldsymbol{\theta}) = \sum_{\{x;s_{(h-1)} \leqslant x < s_h\}} P_J(X = x \mid \boldsymbol{\theta}), \quad (5)$$

其中 $h = 1,2,\cdots,H$.

由此可以计算出能力为 $\boldsymbol{\theta}$ 被试的条件分类一致性指标 $\varphi(\boldsymbol{\theta})$,即 2 个平行测验上能力为 $\boldsymbol{\theta}$ 的被试分类一致的概率为

$$\varphi(\boldsymbol{\theta}) = \sum_{h=1}^{H} (p_{\boldsymbol{\theta}}(h))^2.$$

条件分类一致性指标只是反映固定能力水平的测验分类一致性. 测验对整个能力空间上能力的分类一致性,只需计算 $\varphi(\boldsymbol{\theta})$ 的期望,即可得到测验或边际分类一致性 φ 为

$$\varphi = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} \varphi(\boldsymbol{\theta})\,g(\boldsymbol{\theta})\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_d.$$

为消除随机一致分类偶然概率的影响,Kappa 系数对因随机分类的偶然概率(the chance probability)进行修正,由此可采用下式计算 φ 对应的 Kappa 系数:

$$\kappa = (\varphi - \varphi_c)/(1 - \varphi_c),$$

其中 φ_c 表示由于随机一致分类偶然概率,其计算公式为

$$\varphi_c = \sum_{h=1}^{H} p^2(h).$$

结合(5)式和能力分布,可计算边际分类概率 $p(h)$ 为

$$p(h) = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} p_{\boldsymbol{\theta}}(h)\,g(\boldsymbol{\theta})\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_d.$$

3.1.2 **分类准确性指标** 先计算能力为 $\boldsymbol{\theta}$ 的被试的期望总分或真分数:

$$\tau(\boldsymbol{\theta}) = \sum_{j=1}^{J}\sum_{k=0}^{K_j} kP_{jk}(\boldsymbol{\theta}). \tag{6}$$

设真分数量尺上划界分数为 $\tau_0,\tau_1,\cdots,\tau_H$,其中 $\tau_0 = 0$,划界分数将被试分为 H 类. 根据划界分数,确定能力为 $\boldsymbol{\theta}$ 被试的"真实"类,即当被试真分数满足 $\tau(\boldsymbol{\theta}) \in [\tau_h,\tau_{h+1})$ 时,第 h 类视为被试的"真实"类. 再计算能力为 $\boldsymbol{\theta}$ 被试的条件分类准确性指标 $\gamma(\boldsymbol{\theta})$,即能力为 $\boldsymbol{\theta}$ 的被试分到其"真实"类的概率

$$\gamma(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(h),\text{若 } \tau(\boldsymbol{\theta}) \in [\tau_h,\tau_{h+1}).$$

条件分类准确性指标只是反映给定能力处的测验分类准确性. 测验对整个能力空间上能力的分类准确性,只需计算 $\gamma(\boldsymbol{\theta})$ 的期望,即可得到测验或边际分类准确性 γ 为

$$\gamma = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\gamma(\boldsymbol{\theta})\,g(\boldsymbol{\theta})\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_d.$$

分类准确性指标 γ 对应的 Kappa 系数为

$$\kappa = (\gamma - \gamma_c)/(1 - \gamma_c),$$

其中 $\gamma_c = \sum_{h=1}^{H}\left(\sum_i p_{\boldsymbol{\theta}_i}(h)/N\right)\left(\sum_i w_{ih}/N\right)$,若 $\tau(\boldsymbol{\theta}_i) \in [\tau_h,\tau_{h+1})$,则 $w_{ih} = 1$,否则 $w_{ih} = 0$.

还可以定义条件假阳性率(the conditional false positive error rate)或高估概率、条件假阴性率(the conditional false negative error rate)或低估概率分别为

$$\gamma^+(\boldsymbol{\theta}) = \sum_{h'=h+1}^{H} p_{\boldsymbol{\theta}}(h'),\text{若 } \tau(\boldsymbol{\theta}) \in [\tau_h,\tau_{h+1}),$$

$$\gamma^-(\boldsymbol{\theta}) = \sum_{h'=1}^{h-1} p_{\boldsymbol{\theta}}(h'),\text{若 } \tau(\boldsymbol{\theta}) \in [\tau_h,\tau_{h+1}).$$

边际假阳性率 γ^+ 和边际假阴性率 γ^- 分别为

$$\gamma^+ = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\gamma^+(\boldsymbol{\theta})\,g(\boldsymbol{\theta})\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_d, \tag{7}$$

$$\gamma^- = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\gamma^-(\boldsymbol{\theta})\,g(\boldsymbol{\theta})\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_d. \tag{8}$$

3.2　基于文献[8]方法的分类一致性和分类准确性指标

决策规则是将整个能力空间划分为多个互不相交区域的函数. 若将 d 维能力向量空间 $\mathbf{R}^d$ 划分为 H 个互不相交的决策区域,分别记为 $R_1,R_2,\cdots,R_H$,这 H 个决策区域对应 H 个不同的表现水平.

3.2.1　分类一致性指标　文献[8]的方法是由似然函数计算分类一致的概率. 给定被试 i 的得分 y_i、项目参数估计 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$,由似然函数和决策区域可计算被试 i 被分到第 h 类表现水平的概率为

$$p_{ih} = p_i(R_h) = \frac{\int_{R_h} L(y_i\mid\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\theta}}{\sum_{h=1}^{H}\int_{R_h} L(y_i\mid\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\theta}},$$

其中 $h = 1,2,\cdots,H$,似然函数 $L(y_i\mid\boldsymbol{\theta},\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)$ 见(2)式.

分类一致性为平行测验下各个表现水平上所有被试被分到相同类的比率,即分类一致性 φ 为

$$\varphi = \sum_i\sum_h p_{ih}^2/N.$$

分类一致性 φ 对应的 Kappa 系数为

$$\kappa = (\varphi - \varphi_c)/(1 - \varphi_c),$$

其中 $\varphi_c = \sum_{h=1}^{H} p_h^2 = \sum_{h=1}^{H}\left(\sum_i p_{ih}/N\right)^2$.

3.2.2　分类准确性指标　下面定义基于文献[8]方法的分类准确性指标. 矩阵 $\boldsymbol{W} = (w_{ih})_{N\times H}$ 用于标识被试的表现水平的估计. 如果使用真分数量尺上划界分数,根据被试能力的极大似然估计 $\boldsymbol{\theta}$,由(6)式可计算被试的期望总分或真分数 $\tau(\boldsymbol{\theta})$,再根据划界分数,确定能力为 $\boldsymbol{\theta}$ 的被试的"真实"类. 当被试期望总分满足 $\tau(\boldsymbol{\theta}) \in [\tau_h,\tau_{h+1})$ 时,记 $w_{ih} = 1$,否则 $w_{ih} = 0$. w_{ih} 指示被试的"真实"类. 若使用潜在能力量尺上的决策规则,则可根据被试的能力估计确定 w_{ih}. 由于第 h 类可视为被试 i 的"真实"分类,p_{ih} 即表示被试 i 被分到第 h 类的期望正确分类概率,则正确分类概率或分类准确性指标 γ 为

$$\gamma = \sum_{i=1}^{N}\sum_{h=1}^{H}(p_{ih}w_{ih})/N.$$

分类准确性指标 γ 对应的 Kappa 系数为

$$\kappa = (\gamma - \gamma_c)/(1 - \gamma_c),$$

其中 $\gamma_c = \sum_{h=1}^{H}(p_h w_h) = \sum_{h=1}^{H}\left(\sum_i p_{ih}/N\right)\left(\sum_i w_{ih}/N\right)$.

类似于(7)式和(8)式,边际假阳性率 γ^+ 和边际假阴性率 γ^- 分别为

$$\gamma^+ = \sum_{i=1}^{N}\sum_{h=1}^{h_i-1} p_{ih}/N,$$

$$\gamma^- = \sum_{i=1}^{N}\sum_{h=h_i+1}^{H} p_{ih}/N,$$

其中 $h_i = \arg\max_h(w_{ih})$ 表示被试 i 的"真实"分类.

3.3　基于 Rudner 方法的分类一致性和分类准确性指标

在多维项目反应理论模型下,测验信息量可用于评价能力估计的误差. 例如,能力向量极大似然估计的渐近协方差阵是信息量矩阵的逆矩阵[39]. 多维项目反应理论模型下项目信息量矩阵[40-41]定义如下:

$$\boldsymbol{I}_j(\boldsymbol{\theta}) = -E\left(\partial^2\log L(Y_j\mid\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}\right),$$

其中 $L(Y_j\mid\boldsymbol{\theta})$ 表示项目 j 上的似然函数,可由(2)式变化而来. 对于多维等级反应模型下项目信息量矩阵 $\boldsymbol{I}_j(\boldsymbol{\theta})$ 主对角线元素计算公式如下:

$$(\boldsymbol{I}_j(\boldsymbol{\theta}))_{ll} = \sum_{k=0}^{K_j}\frac{1}{P_{jk}(\boldsymbol{\theta})}\left(\frac{\partial P_{jk}(\boldsymbol{\theta})}{\partial\theta_l}\right)^2 =$$

$\sum_{k=0}^{K_j}(a_l P_{jk}^*(\boldsymbol{\theta}_i)(1-P_{jk}^*(\boldsymbol{\theta}_i)) - a_l P_{jk+1}^*(\boldsymbol{\theta}_i)(1-P_{jk+1}^*(\boldsymbol{\theta}_i)))^2/(P_{jk}^*(\boldsymbol{\theta}_i)-P_{jk+1}^*(\boldsymbol{\theta}_i))$,

其中 $l=1,2,\cdots,d$. 单维模型下信息量计算公式的可参见相关文献[22,42-43]. 项目信息量矩阵 $\boldsymbol{I}_j(\boldsymbol{\theta})$ 非主对角线元素计算公式如下:

$$(\boldsymbol{I}_j(\boldsymbol{\theta}))_{ll'}=\sum_{k=0}^{K_j}\frac{1}{P_{jk}(\boldsymbol{\theta})}\left(\frac{\partial P_{jk}(\boldsymbol{\theta})}{\partial\theta_l}\right)\left(\frac{\partial P_{jk}(\boldsymbol{\theta})}{\partial\theta_{l'}}\right),$$

其中 $l,l'=1,2,\cdots,d,l\neq l'$. 项目信息量矩阵 $\boldsymbol{I}_j(\boldsymbol{\theta})$ 的公式如下:

$$\boldsymbol{I}_j(\boldsymbol{\theta})=\{\sum_{k=0}^{K_j}(P_{jk}^*(\boldsymbol{\theta}_i)(1-P_{jk}^*(\boldsymbol{\theta}_i))-P_{jk+1}^*(\boldsymbol{\theta}_i)(1-P_{jk+1}^*(\boldsymbol{\theta}_i)))^2/(P_{jk}^*(\boldsymbol{\theta}_i)-P_{jk+1}^*(\boldsymbol{\theta}_i))\}\cdot\begin{bmatrix}a_{j1}^2 & a_{j1}a_{j2} & \cdots & a_{j1}a_{jd}\\ a_{j2}a_{j1} & a_{j2}^2 & \cdots & a_{j2}a_{jd}\\ \cdots & \cdots & \cdots & \cdots\\ a_{jd}a_{j1} & a_{jd}a_{j2} & \cdots & a_{jd}^2\end{bmatrix}.$$

在局部独立假设条件下,项目信息量具有可加性[42],由此得到能力点 $\boldsymbol{\theta}$ 处的测验信息量矩阵为

$$\boldsymbol{I}^{(J)}(\boldsymbol{\theta})=\sum_{j=1}^{J}\boldsymbol{I}_j(\boldsymbol{\theta}).$$

下面介绍基于信息量矩阵的分类一致性和分类准确性指标. 能力向量的极大似然估计渐近服从多元正态分布,记为 $\boldsymbol{\theta}\sim M_{VN}(\hat{\boldsymbol{\theta}}_i,\boldsymbol{I}^{(J)}(\hat{\boldsymbol{\theta}}_i))$[39,42,44]. 由多元正态分布可计算被试 i 分到第 h 类的期望概率为

$$\hat{p}_{ih}=\int_{R_h}\frac{1}{(2\pi)^{d/2}|\boldsymbol{I}^{(m)}(\hat{\boldsymbol{\theta}}_i)|^{-1/2}}\exp(-(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_i)^{\mathrm{T}}\boldsymbol{I}^{(J)}(\hat{\boldsymbol{\theta}}_i)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_i)/2)\,\mathrm{d}\boldsymbol{\theta},$$

其中 $|\boldsymbol{I}^{(J)}(\hat{\boldsymbol{\theta}}_i)|$ 表示能力点 $\hat{\boldsymbol{\theta}}_i$ 处的测验信息量矩阵. 该积分式可通过数值积分方法的蒙特卡罗模拟方法计算. 由此,可计算分类一致性 $\hat{\varphi}$ 和分类准确性指标 $\hat{\gamma}$ 分别如下:

$$\hat{\varphi}=\sum_{i=1}^{N}\sum_{h=1}^{H}\hat{p}_{ih}^2/N,\hat{\gamma}=\sum_{i=1}^{N}\sum_{h=1}^{H}(\hat{p}_{ih}\hat{w}_{ih})/N.$$

基于 Rudner 方法的分类一致性 $\hat{\varphi}$ 和分类准确性指标 $\hat{\gamma}$ 对应的 Kappa 系数,可类似于文献[8]的 Kappa 系数计算.

4 分类一致性和分类准确性的价值

CRT 根据测验分数和决策规则只将被试在各个维度掌握程度上分成少数几类表现水平. 因为分类的类数少,在各个内容维度只需较少试题便可得到较好的分类精度,特别适合于大规模测评等. 前已述及,许多大型 CRT 具有多维性. 若不同能力维度之间存在相关性,则由于 MIRT 可以互借不同维度信息从而提高分类结果的信度和效度,因此,MIRT 是分析多维测验数据的重要方法之一. 众多研究者介绍了分类一致性和分类准确性指标及其估计方法,有必要分析其应用条件、应用场合及其价值.

这些指标可用于估计单个测验的分类一致性和分类准确性. 无需进行重复测量,也无需采用能力分布和项目参数估计模拟平行测验再估计分类一致性和分类准确性. 测验的分类一致性尽管可以通过重复测量计算,但重复测量条件比较苛刻,在实际应用中较难获得重测数据[11]. 而对于测验的分类准确性,在真实测验情景下被试的真实能力未知,无法计算估计能力与被试真实能力分类相同的比率.

能力分数或观察总分的条件标准误差[1]也可用来评价 CRT 的分类误差,但是它并不能直接等同于测验的分类准确性. 条件标准误差反映能力估计值与能力"真值"之间的渐近误差大小,在测验长度较短时可能未必合适. 条件标准误差可反映测验在各个能力处的标准误差,并未直接显示测验的整体分类准确率. 不过,当单维 IRT 模型能力误差分布为正态分布时,条件标准误差与测验分类准确性存在非线性转换关系[45]. 在多元正态分布假设下,理论上这种关系在 MIRT 模型中很可能仍然成立,但有待深入研究.

本文介绍的指标及估计方法可用于模拟研究和实证研究. 只需在调用 MIRT 模型的参数估计程序之后再调用指标估计的实现算法,就可基于测验作答数据、项目参数估计、估计的能力分布和决策规则(或划界分数),也可计算或估计真实测验的分类结果的分类一致性和分类准确性指标,用于反映分类结果的信度和效度. 另外,基于观察分数量尺的分类一致性和分类准确性指标已经用于评价真实测验的分类信度和效度. 例如,在单维 IRT 模型或其他统计模型下,已有研究[10]表明文献[7]的方法已经用于评价许多真实测验的分类结果质量,并且已经开发了专门的商业或免费软件供用户使用.

这些方法或指标可用于评价复杂决策规则和多维模型下域分数(domain scores)或子分数(subscores)的质量. 域分数或子分数可反映被试对某个内容、知识或技能的掌握程度,它比量表分数解释性

更好，大众接受度也更高[46]. 因为 IRT 或 MIRT 具有参数不变性和成熟的等值方法、可以利用维度间信息相关从而提高各个子分数的分类信度和效度等优势，基于 IRT 或 MIRT 模型的领域分数或子分数更具优势.

5　结论

本研究介绍了 MGRM 下的分类一致性和准确性指标，下面对已有研究的相关结论进行归纳和总结: 3 类方法均可较好地用于多维模型下的分类一致性和准确性估计，可用于多维 CRT 表现水平的信度和效度评价; 类似于单维模型的结论，在多维模型下，基于潜在能力量尺分数的 2 类方法(文献 [8] 方法和文献 [13] 方法) 比基于观察分数的方法(文献 [7] 方法) 所得到的分类一致性略高，在能力之间相关性较大时分类准确性更高; 3 类方法中涉及的求和或积分可通过蒙特卡罗模拟方法估计; 基于潜在能力量尺的 2 类方法比基于观察分数量尺的方法应用范围更广，可适用于多种决策规则指标估计(既适合于能力分数指标估计，还适合于内容或技能子分数、合成分数等指标估计) ; 在总分决策规则和无信息先验分布下(即先验分布为均匀分布) ，文献 [7-8] 方法下分类准确性指标估计量依概率收敛于同一真值.

6　讨论

不同于 Rudner 的方法[12-13]，文献 [8] 方法可适用于非正态性数据，无需借助能力估计误差渐近正态性假设[8]，这样可避免分数正态转换过程可能引起分类结果差异的问题[5]. 测验长度越长，极大似然法估计的渐近正态性满足越好. 已有研究并没有考虑在能力估计误差分布为非正态分布条件下各指标的表现. 当能力估计误差分布为非正态分布时，各指标尤其是 Rudner 指标的稳健性如何，有待研究. 在不同条件下，有待将本文介绍的指标估计方法与非参数估计方法[10]进行比较.

因为各指标的估计方法均依赖于测量模型，在实际应用中不能单纯考虑分类一致性和分类准确性的高低，还需要考虑模型-资料拟合等其他信度和效度的影响因素. 例如，文献 [8] 方法需要基于项目反应函数计算似然函数; 文献 [13] 方法需要利用能力估计的信息矩阵，信息矩阵同样依赖于似然函数; 文献 [7] 方法也同样依赖于似然函数或联合概率分布. 另外，能力向量的信息矩阵还可以采用不同的估计方法得到，信息矩阵的不同估计方法对指标估计的影响如何，也有待考虑.

若以合成能力分数信息量最大或分类准确性最高为目标求取分数合成的权重[47]，则不等权重的合成分数是否可显著提高分类结果的分类一致性和准确性值得探讨. 在特定应用中，需要综合考虑测验目的、结构效度、内容效度、分数解释性、测验公平性和决策风险等因素决定决策规则. 对于计算机分类测验，分类一致性和分类准确性指标在计算机自动组卷、计算机多阶段自适应测验构建中的应用，也需要探讨.

7　参考文献

[1] 戴海琦. 心理测量学 [M]. 北京: 高等教育出版社，2010.

[2] 甘良梅，余嘉元. 标准参照测验分数体系的探讨研究 [J]. 心理学探新，2006，26(3) : 79-83.

[3] 辛涛，李勉，任晓琼. 基础教育质量监测报告撰写与结果应用 [M]. 北京: 北京师范大学出版集团，2015.

[4] Duncan A. Address by the secretary of education at the 2009 governors education symposium: states will lead the way towards reform [EB/OL]. http: //www2. ed. gov/news/speeches/2009/06/06142009. pdf.

[5] Douglas K M，Mislevy R J. Estimating classification accuracy for complex decision rules based on multiple scores [J]. Journal of Educational and Behavioral Statistics，2010，35(3) : 280-306.

[6] 陈平，李珍，辛涛，等. 标准参照测验决策一致性指标研究的总结与展望 [J]. 心理发展与教育，2011(2) : 210-215.

[7] Lee W C，Brennan R L，Wan L. Classification consistency and accuracy for complex assessments under the compound multinomial model [J]. Applied Psychological Measurement，2009，33(5) : 374-390.

[8] Guo Fanmin. Expected classification accuracy using the latent distribution [J]. Practical Assessment，Research and Evaluation，2006，11(6) : 1-6.

[9] Lathrop Q N，Cheng Ying. Two approaches to estimation of classification accuracy rate under item response theory [J]. Applied Psychological Measurement，2013，37(3) : 226-241.

[10] Lathrop Q N，Cheng Ying. A nonparametric approach to estimate classification accuracy and consistency [J]. Journal of Educational Measurement，2014，51(3) : 318-334.

[11] Lee W C. Classification consistency and accuracy for complex assessments using item response theory [J]. Journal of Educational Measurement,2010,47(1):1-17.

[12] Wyse A E,Hao Shiqi. An evaluation of item response theory classification accuracy and consistency indices [J]. Applied Psychological Measurement,2012,36(7):602-624.

[13] Rudner L M. Expected classification accuracy [J]. Practical Assessment Research and Evaluation,2005,10(13):1-4.

[14] Yao Lihua. Classification accuracy and consistency indices for summed scores enhanced using mirt for test of mixed item types [EB/OL]. [2018-12-16]. http://www.bmirt.com/8220.html.

[15] LaFond L J. Decision consistency and accuracy indices for the bifactor and testlet response theory models detecting heterogeneity in logistic regression models [EB/OL]. [2018-12-21]. https://ir.uiowa.edu/etd/1346.

[16] Debeer D,Buchholz J,Hartig J,et al. Student,school,and country differences in sustained test-taking effort in the 2009 pisa reading assessment [J]. Journal of Educational and Behavioral Statistics,2014,39(6):502-523.

[17] Makransky G,Mortensen E L,Glas C A W. Improving personality eacet scores with multidimensional computer adaptive testing: an illustration with the Neo Pi-R [J]. Assessment,2012,20(1):3-13.

[18] Rijmen F,Jeon M,von Davier M,et al. A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys [J]. Journal of Educational and Behavioral Statistics,2014,39(4):235-256.

[19] Yao Lihua,Boughton K A. A multidimensional item response modeling approach for improving subscale proficiency estimation and classification [J]. Applied Psychological Measurement,2007,31(2):1-23.

[20] Zhang Jinming. Calibration of response data using MIRT models with simple and mixed structures [J]. Applied Psychological Measurement,2012,36(5):375-398.

[21] Cai Li. High-dimensional exploratory item factor analysis by a metropolis-hastings robbins-monro algorithm [J]. Psychometrika,2010,75(1):33-57.

[22] Reckase M D. Multidimensional item response theory [M]. New York: Springer,2009.

[23] 刘红云,骆方,王玥,等. 多维测验项目参数的估计:基于 SEM 与 MIRT 方法的比较 [J]. 心理学报,2012,44(11):121-132.

[24] 杜文久,肖涵敏. 多维项目反应理论等级反应模型 [J]. 心理学报,2012,44(10):1402-1407.

[25] 康春花,辛涛. 测验理论的新发展:多维项目反应理论 [J]. 心理科学进展,2010,18(3):530-536.

[26] 涂冬波,蔡艳,戴海琦,等. 多维项目反应理论:参数估计及其在心理测验中的应用 [J]. 心理学报,2011,43(11):1329-1340.

[27] 许志勇,丁树良,钟君. 高考数学试卷多维项目反应理论的分析及应用 [J]. 心理学探新,2013,33(5):438-443.

[28] 詹沛达,王文中,王立君,等. 多维题组效应 Rasch 模型 [J]. 心理学报,2014,46(8):1208-1222.

[29] 汪文义,宋丽红,丁树良. 复杂决策规则下 MIRT 的分类准确性和分类一致性 [J]. 心理学报,2016,48(12):1612-1624.

[30] Wang Wenyi,Song Lihong,Ding Shuliang,et al. Estimating classification accuracy and consistency indices for multidimensional latent ability [EB/OL]. [2018-10-12]. https://link.spriger.com/chapter/10.1007%2F978-3-319-38759-8-8.

[31] Wang Wenyi,Song Lihong,Ding Shuliang. An extension of rudner-based consistency and accuracy indices for multidimensional item response theory [EB/POL]. [2018-12-11]. www.doc88.com/p-3149195293902.html.

[32] Chalmers R P. MIRT: a multidimensional item response theory package for the r environment [J]. Journal of Statistical Software,2012,48(6):1-29.

[33] Henderson-Montero D,Julian M W,Yen W M. Multiple measures alternative design and analysis models [J]. Educational Measurement: Issues and Practice,2003,22(2):7-12.

[34] Chester M D. Multiple measures and high-stakes decisions a framework for combining measures [J]. Educational Measurement: Issues and Practice,2003,22(2):32-41.

[35] McBee M T,Peters S J,Waterman C. Combining scores in multiple-criteria assessment systems: the impact of combination rule [J]. Gifted Child Quarterly,2014,58(1):69-89.

[36] Carroll P E,Bailey A L. Do decision rules matter? A descriptive study of english language proficiency assessment classifications for english-language learners and native english speakers in fifth grade [J]. Language Testing,2016,33(1):23-52.

[37] Abedi J. The no child left behind act and english language learners: assessment and accountability issues [J]. Educational Researcher,2004,33(1):4-14.

[38] Chang Huahua. Making computerized adaptive testing diagnostic tools for schools [C]// Lissitz R W,Hong Jiao. Computers and their impact on state assessment: recent history and predictions for the future. Charlotte,NC: Information Age Publisher Inc,2012:195-226.

[39] Wang Chun. On latent trait estimation in multidimensional

compensatory item response models [J]. Psychometrika, 2015,80(2):428-449.

[40] Ackerman T A. Full-information factor analysis for polytomous item responses [J]. Applied Psychological Measurement,1994,18(3):257-275.

[41] Yao Lihua,Schwarz R D. A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests [J]. Applied Psychological Measurement,2006,30(6):469-492.

[42] Chang Huahua. The asymptotic posterior normality of the latent trait for polytomous irt models [J]. Psychometrika, 1996,61(3):445-463.

[43] Samejima F. Estimation of latent ability using a response pattern of graded scores [J]. Psychometrika, 1969, 34(1):1-97.

[44] Chang Huahua,Stout W. The asymptotic posterior normality of the latent trait in an irt model [J]. Psychometrika, 1993,58(1):37-52.

[45] Cheng Ying,Liu Cheng,Behrens J. Standard error of ability estimates and the classification accuracy and consistency of binary decisions [J]. Psychometrika,2015,80(3):645-664.

[46] 辛涛,谢敏. 群体水平领域分数及其估计方法 [J]. 心理发展与教育,2010(4):416-422.

[47] Yao Lihua. Multidimensional linking for domain scores and overall scores for nonequivalent groups [J]. Applied Psychological Measurement,2010,35(1):48-66.

The Quality Evaluation Index for Score Reporting in Multidimensional Criterion-Referenced Tests

SONG Lihong[1], WANG Wenyi[2]

(1. Elementary Education College, Jiangxi Normal University, Nanchang Jiangxi 330022, China;
2. College of Computer Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: For criterion-referenced tests, classification consistency and accuracy are important indicators for evaluating the reliability and validity of classification results in scores reporting. Numerous procedures have been proposed to estimate these indices in the framework of unidimensional item response theory (UIRT). Multidimensional item response theory (MIRT) has been devoted to models that include more than one latent trait to account for the multidimensional nature of complex constructs. MIRT has been successfully employed to analyze many criterion-referenced tests. Because MIRT has enjoyed tremendous growth, the purpose of this study will give a brief review of decision rules and three types of classification consistency and accuracy. The first one is the classification accuracy and consistency based on total sum scores, the second is the likelihood-based consistency and accuracy, and the last is the information-based consistency and accuracy. Finally, two practical implications of this research have been identified. First, it is easily to estimate classification consistency and accuracy indices for subscores or composite scores in each knowledge, content or skill area when the true cut scores were on the total score or latent ability scale. Second, they might be useful for developing test construction method in a multistage testing which is a form of computerized adaptive classification testing for making classification decisions.

Key words: multidimensional item response theory; score reporting; decision rule; classification accuracy; classification consistency

(责任编辑: 冉小晓)