http://myfj.cnjournals.com myfj_sadri@comac.cc (021)20866796

DOI: 10.19416/j.cnki.1674-9804.2022.03.022

基于 Rasch 数学模型的飞行理论测评方法

王 真*

(东航技术应用研发中心有限公司,上海 201707)

摘 要: 培养合格的飞行员无论对航空公司还是对乘客都至关重要。为了符合航空法规要求和飞行安全,航空公司要定期对所有飞行员进行培训及评测。研究如何通过测试更好地评价飞行员、测评分数,明确能够真实客观反映飞行员水平的能力指标,对航空公司及培训机构都有重大意义。在基于项目反应理论(Item Response Theory,简称 IRT)的 Rasch 数学模型的基础上,通过分析计算试题的难度与质量、试题的鉴别度、人的分数能力等重要参数,建立一套自适应的理论测评方法,进而自动判断试卷的可靠性,同时用真实的测评数据进行回归分析,从而建立一套独特的自适应测评管理方法。在全球范围内首次提出飞行员胜任力三层理论这个重要概念,为飞行员能力评价及试卷可靠性评价提供了可行的实施方案和重要理论依据。

关键词: IRT 模型;Rasch 模型;项目反应理论;测试分析模块包;最大似然函数

中图分类号: V211.1; TP391.91 **文献标识码:** A **OSID:**

0 引言

尽管理论水平并不能完全体现飞行员的飞行能力,但理论是飞行实践的基础,尤其是对于民航飞行员来说,现代飞机的复杂性以及各界对于安全的高度敏感性要求飞行员具备全面且扎实的飞行理论。大西洋上空 A330 空中失速坠毁一案充分说明飞行员需要良好的气动原理的知识才可能正确应对空中失控这样复杂的情况。波音 737MAX 两起飞行事故也说明飞行员如果不清楚飞机系统的工作原理,将很难处理飞机故障引发的特情。为帮助飞行员提升理论知识的掌握程度,航空公司需要一套科学合理的理论评估方法,能够准确可观评估飞行员的理论水平,并有针对性地持续开展培训。

1 IRT 算法背景

项目反应理论(Item Response Theory,简称 IRT)是一套基于潜在变量的心理测量理论,专门设计用来模拟考生“能力”和项目水平刺激(难度、猜测等)之间的交互作用。其重点是考生对测试题的反应模式,而非复合或总分变量及线性回归理论。IRT 框架强调从概率角度考虑响应。在 IRT 中,考生对测试项目的回答被认为是结果(因变量),而考生能力和项目特征是潜在的预测(独立)变量。

古典测验理论(Classical Test Theory,简称 CTT)兴起于十九世纪末,很长时间内占据着心理测量学的统治地位。直到 1953 年,Lord[1-2] 发表了有关潜伏特质理论的博士学位论文,IRT 才开始成为主流方法。CTT 根据真实分数和观察分数之间的线性关系(观察分数 = 真实分数 + 错误)对测试结果进行建模,而 IRT 根据人的能力和所测项目的特征对考生的响应模式概率进行建模、测试或调查。

经过几十年的发展,IRT 理论衍生出大量测量模型。Rasch 在其著作中提出了项目响应模型 Rasch model[3]。Samejima 在 1969 年提出了分级反应模型[4]。1970 年到 2020 年间,一批新的学者在这一领域投入了研究热情,包括 Hambleton 和 Sawminathan[5],Fox[6],Reckase[7],Rijmen 和 Tuerlinckx[8],Jabarayilov、Emons 和 Sijtsma[9],Fisher 和 Molenaar[10],Gin[11],Marquardt 和 Pemstein[12],Camilli 和 Geis[13]。2007 年 Mair 和 Hatzinger 提出 Rasch 扩展模型并给出 R 语言的现实方法[14]。2017 年和 2019 年,Swartz

* 通信作者. E-mail: 2316606698@qq.com

引用格式: 王真. 基于 Rasch 数学模型的飞行理论测评方法[J]. 民用飞机设计与研究,2022(3):133-140. WANG Z. Flight theory evaluation method based on Rasch model[J]. Civil Aircraft Design and Research,2022(3):133-140(in Chinese).

和 Guan 提出一个基于 Bayesian 理论的新的 IRT 算法[15-16]。2020 年李森森把 IRT 理论应用到飞行员理论测评[17],这些学者在这一领域相继作出重大贡献。

2 航空理论测评背景

目前航空公司在飞行理论培训方面存在以下几个挑战:

1)理论知识覆盖面不够全面,知识盲点多。比如前面所讲飞行原理知识,容易被航空公司忽略;

2)培训内容针对性不足。每年几乎都是机型手册、除防冰一类,没有充分利用大数据的结果获得机队整体知识结构然后进行针对性培训;

3)每个飞行员几乎得到相同的培训与考核。这样的培训与 EBT 比较既浪费时间效果又不佳,这同样是因为没有利用大数据的支持;

4)考核形式陈旧。目前普遍采用的 80 分制不能矫正由于题目难度和效度导致的评价偏差。同时因为不是标准分,组织和个体也很难知晓个体成绩在团体数据中的位置,如果采用标准分则可能规定两个标准差之外的考生需要再评估;

5)对于题目本身缺少评价。题目的难易度、效度或鉴别度等没有被考虑,而有些题目可能本身是有歧义的甚至是错的,需要被统计发现出来;

6)考试题库更新不够及时,无法满足实际运行对飞行员能力的需求。试题难度和准确度没有统一的衡量标准,导致试题质量参差不齐,进而使测试效果无法达到预期。

因此航空公司需要新设计一套基于大数据的、能够紧贴实际运行的、高品质、高标准反映飞行员能力的理论评估系统。首先,该系统需要具备完整的试题库,涵盖飞行员所必需的所有知识范围;其次该系统应能对题目持续进行评价;此外,应能输出个体和团体的成绩数据分布,航空公司可以据此对团体和个体实施有针对性的培训,并且能对飞行员在理论知识维度进行“数字化画像”。Rasch 算法为上述需求提供了很好的解决方案。

3 Rasch Model 原理

3.1 多模 Rasch 模型

Rasch 在其关于基本 Rasch 模型的第一个出版物问世之后,开始研究 Rasch 的多态概括。1995 年,Andersen 得出了以下基于 Rasch 模型对多数据的一般表达。数据矩阵表示为 X,行表示人员,列表示项目。X 中的单个元素表示为 x_{vi},其中 $v = 1,\ldots,n$(共有 n 个人),而 $i = 1,\ldots,k$(共有 k 个项目)。此外,每个项目 i 具有一定数量的响应类别,用 $h = 0,\ldots,m_i$ 表示。可以根据以下两个表达式得出项目 i 上相应的响应 h 的概率:

$$P(X_{vi} = h) = \frac{e^{\phi_h(\theta_v+\beta_i)+\omega_h}}{\sum_{l=0}^{m_i} e^{\phi_l(\theta_v+\beta_i)+\omega_l}} \tag{1}$$

或

$$P(X_{vi} = h) = \frac{e^{\phi_h\theta_v+\beta_{ih}}}{\sum_{l=0}^{m_i} e^{\phi_l\theta_v+\beta_{il}}} \tag{2}$$

式(1)中,ϕ_h 为项目参数的评分函数;θ_v 为关于人的参数;β_i 为项目参数;ω_h 为类别参数。式(2)中,β_{ih} 为项目类别参数。在这两个公式的框架内,许多模型都建议保留 Rasch 模型的基本属性,以便应用基于条件最大似然估计(Conditional Maximum Likelihood,简称 CML)。

3.2 简化 Rasch 模型

当得知考试项目的难度(β_i)和被测试者的能力(θ_v)时,式(1)的模型可简化为:

$$P(X_{vi} = 1) = \frac{e^{\theta_v-\beta_i}}{1 + e^{\theta_v-\beta_i}} \tag{3}$$

式(3)中,P 为测试者答对题目的概率。这个模型适用的主要假设是:潜在特征的一维性,原始分数的充分性,局部独立性和平行项特征曲线(Item Characteristic Curve,简称 ICC)。相应的解释由 Fischer 给出,并进行了数学推导和证明。

4 测评方法

培养飞行员胜任力的目的是让飞行员在实际飞行运行中通过胜任力的完全展示实现安全运行。本文以 Rasch 模型和 Dr. Swartz 的理论为主要依据,吸收借鉴了“Kirkpatrick 四层训练评估模型”的相关理论,结合航空公司的实际情况,首先提出了一套针对飞行训练中胜任力训练和评估的模型——飞行胜任力三层理论模型,如图 1 所示。

该模型中,每一层胜任力的获得都包含两个部分:教学和评估。评估的主要目的是确认受训者在完成训练之后的胜任力水平是否满足训练预期。评估环节包括三个部分:问卷调查、课内评估和课后评

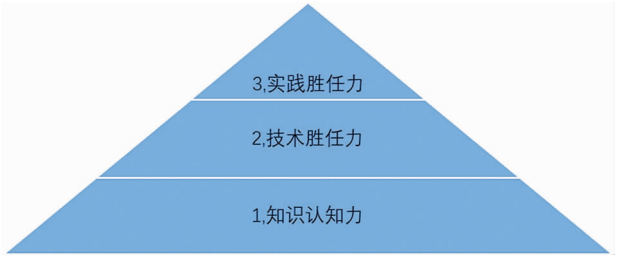


图 1　飞行胜任力三层理论示意图

估。问卷调查是为了收集受训者对训练的反馈,以便训练模型的改进和调整。课内评估的目的是对受训者刚刚习得的知识或技能的表现做评估,并作为该层初始胜任力记录。课后评估,是在训练结束之后对受训者进行的综合性评估。根据训练者所处训练层级的不同,胜任力的评估内容、评估方式和评估标准也不相同。本文研究对象仅限于模型的第一层级:知识认知力。

该能力的评估需依靠一套完善的自适应考试系统,该系统试题库必须涵盖足够的航空知识点而且能够有效地定位飞行员在每个知识点对应的认知力。

4.1 试题库的特点

为分析飞行员综合能力和各种试题的相关指标,试题库必须具有大体量、多样性、真实性和充分性等特点。

大体量:以中国四大航平均飞行员个数 6 000 为例,试题应设置为三倍飞行员个数,即 18 000 道题。试题量越大,检验的效度就越大。

客观性:试题全部以客观题选择题形式呈现,使测试结果更接近真实结果。

多样性:除了对知识记忆进行测试,还会通过场景模拟测试考生对理论知识的实践效果,中间还可以添加计算过程等,以达到多方面检测考生胜任力的目的。

真实性:每一份数据都必须来源于每位飞行员参与考试的真实作答。

充分性:为了确保数据分析的准确性,每道题的作答次数必须至少达到 20 人次,考试系统在组卷抽题时应考虑这一要求。

全面性:每道题必须建立动态的难度和鉴别度属性,考试系统在组卷抽题时应该考虑这些参数。

4.2 建立全面的知识能力体系和完善的评估标准

分类:将知识类型分为五大类:非技术知识、程序、通用航空知识、飞机系统知识、公司运行规章与法规。为了更细致地表征不同类型,总结了十大知识类,包括机型系统,性能,正常和非正常程序,航空气象等,其中每一类又可细分为 3 ~ 15 个不等的知识点,共归纳了 112 个知识点。

过程:将每个知识类型的所有作答记录汇总成一张类型试卷,对该类型试卷进行一次自动推导,计算出考生的各个知识类型的能力。

优劣对比:系统通过整体分析得到飞行员的各个知识类型能力,相比于传统的分数均值,具有不会显著随样本波动而波动,或是随试题整体难度波动而波动的优势。

4.3 更为科学的衡量指标——能力与分数相结合的分析方式

单纯的分数说明不了一个考生在学校的学习效果,更科学的方法是用 IRT 理论进行转换。

优劣对比:系统采取全局试卷分析得到的全局能力分析可以减少偶然因素的影响(主要指试题难度的波动性),从而能更加准确客观地得到一个考生能力的估值。传统的 CTT 是确定性模型,它的优点在于可以快速区别出大多数考生的能力,但是一个考生的得分依赖于试题难度,很难得到一个客观的评价。CTT 与 IRT 的二者结合才能更好地表现一个考生的擅长处并提供一种相对较好的学习方式[17],将能力值和分数结合起来,能更准确地对考生进行评测,多层次、多维度地分析考生的能力提高路径。

4.4 能够自动推导考生能力和试题难度以及试卷鉴别度

传统的 CTT 模型(以分数为基础)会受试题难度的影响,且试题难度的估计也会受考生能力的影响,因此很难得到在考生能力和试题难度这两种因素波动下较为准确的对于考生能力和试题难度的评估。

本文算法基于项目反应理论,将考生能力和试题难度结合起来进行计算,将两者放在同一刻度下进行衡量,计算出较为准确的考生能力和试题难度以及鉴别度,能很好地避免出题者的主观判断导致试卷可靠性不高的情况,并将考生能力和试题难度划分为五个等级,便于参考。

5　方法的应用举例

5.1 考试系统流程(见图 2)

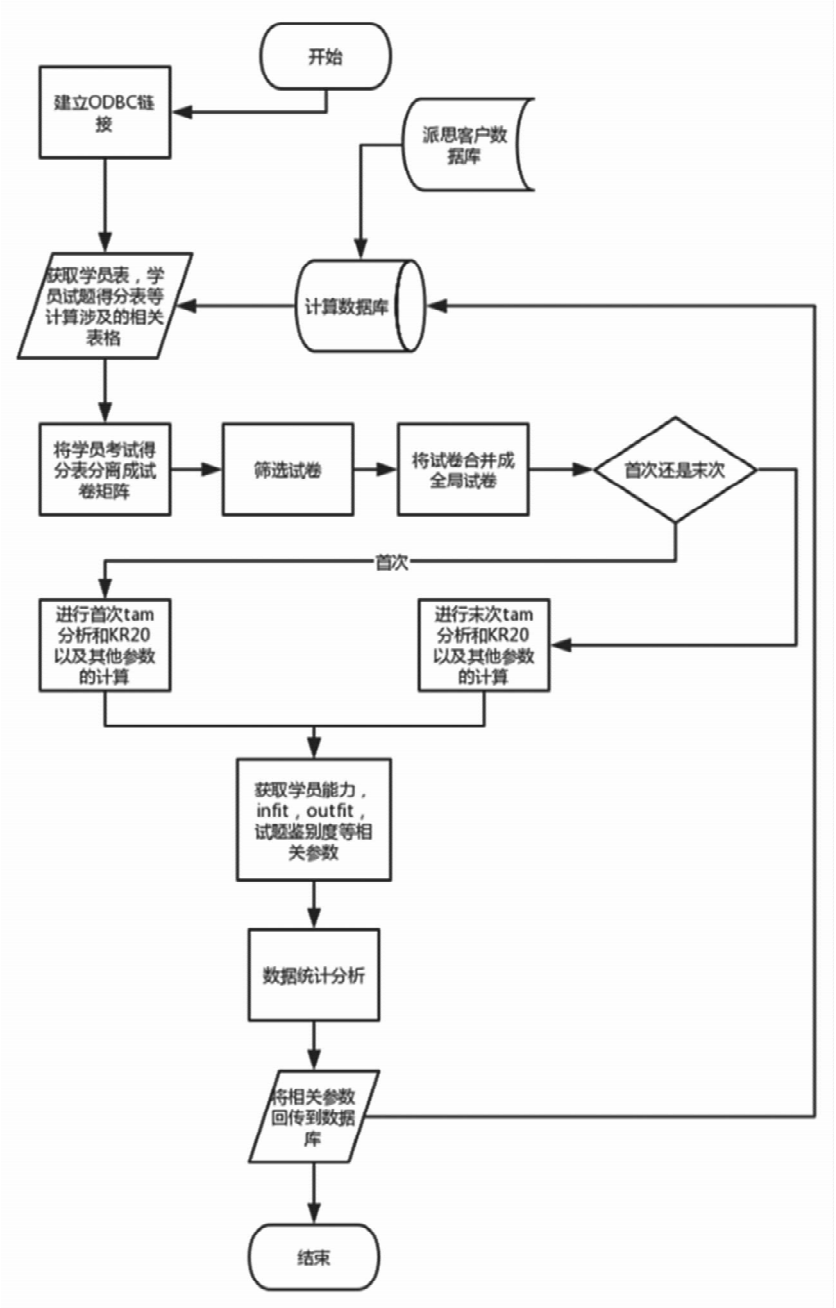


图 2　考试系统自动推导的流程图

5.2 流程详细说明

1）读取真实数据并初始化数据

首先是加载库，取出数据，并把考生试题得分大于等于题目一半的分数设为 1，小于题目一半总分的设为 0。

2）分离试卷并筛选数据

将考生考试得分分离成每张试卷的每个考生对应的每道题的答题结果，同时去掉每一张试卷里的空白答题结果，去掉参与人数小于 5 的试卷答题结果和少于四道试题的试卷。

3）合并试卷

将所有试卷合并成一张全局试卷。

4）单张试卷能力分析

由于 Rasch 模型限制了每列的取值范围，所以首先去掉全列和为 0 的列，然后通过 R 进行能力的分析，并把分析出的能力存储起来，然后再将每次考试每个考生对应的能力合并为一张能力表。

5）试卷可靠度计算

采用如公式(4)所示的公式计算试卷的可靠度:

$$KR20 = \frac{k}{k-1}\left(1 - \frac{\sum pq}{S^2}\right) \tag{4}$$

式(4)中,k 为考生的数量;p,q 为每道题目答对或答错的概率;S^2 为所有考生总分的方差;KR20 为计算出来的试卷信度,一般认为 KR20 大于 0.7 的试卷是一套较好的试卷。

分析结果的主要参数包括:试题的难度、Infit、Outfit、正确率、试题的鉴别度、考生的能力、试题模板 KR20。

5.3 应用结果分析

本文对航空公司提供的两百万条飞行员在航培时的真实作答结果进行了如上方法的数据分析,认为:考生的分数应该与试题难度紧密相关,因为一道题答对的考生越多,说明这道题越容易。因此本文使用简化 Rasch 模型来计算考生的能力值和题目难度,该模型的好处是将考生能力和试题难度统一起来,在同一个刻度下进行衡量。首先按照流程①的做法将考生作答结果初始化为 0(答错)或 1(答对),将考生的最终成绩作为原始能力值 θ_{01},θ_{02},θ_{03},…。设 β_{0i} 是第 i 题的初始难度,则针对第 v 名飞行员对第 i 道试题的作答结果,用 Rasch 模型表示出其回答正确的概率如式(5):

$$P(X_{vi}) = \frac{e^{\theta_{0v}-\beta_{0i}}}{1+e^{\theta_{0v}-\beta_{0i}}} \tag{5}$$

回答错误的概率如式(6):

$$Q(X_{vi}) = 1 - \frac{e^{\theta_{0v}-\beta_{0i}}}{1+e^{\theta_{0v}-\beta_{0i}}} = \frac{1}{1+e^{\theta_{0v}-\beta_{0i}}} \tag{6}$$

对于每一道试题的难度,采用如式(7)所示的将所有概率相乘再取对数,然后用如式(8)所示的求最大似然估计的方法进行计算:

$$L_{0i} = \ln\prod_{v=1}^{V}\left(P(X_{vi})^{ui}Q(X_{vi})^{1-ui}\right) \tag{7}$$

$$\hat{\beta}_{0i} = \operatorname{argmax}(L_{0i}) \tag{8}$$

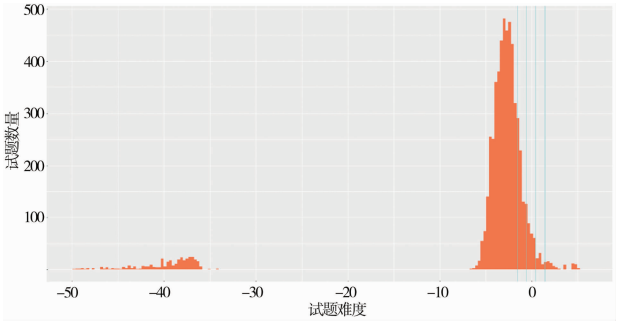
式(7)中,ui 为作答正确或错误,如果正确则 $ui=1$,如果错误则 $ui=0$。

将计算出的每一道题的难度代入 Rasch 模型中,根据该难度和考生的作答结果用同样的最大似然估计方法反过来计算每个飞行员的能力值 θ_{11},θ_{12},θ_{13},…。通过不断重复的迭代计算,当飞行员的能力值和试题的难度收敛至不再发生变动时,就得出最终的能力值和试题难度。

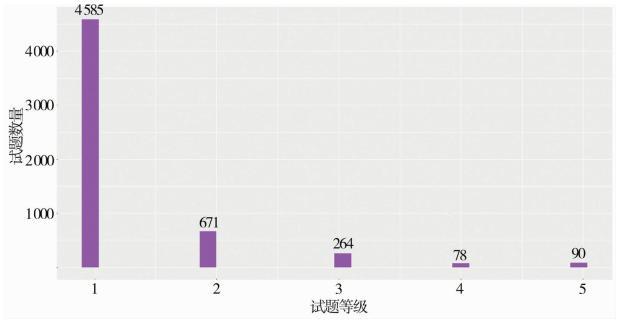
同时,将飞行员能力值和题目难度划分为 5 个等级,1 级为最低等级,5 级为最高等级。不同的航空公司由于数据的不同会得到不一样等级的算法。

图 3(a)是经过本文算法计算出试题难度后,通过可视化生成的直方图。由图可见,题目难度以 -2.613 为中心呈理想的正态分布。

图 3(b)是将题目数量按难度分成 5 级后的柱状图。



(a) 试题难度分布

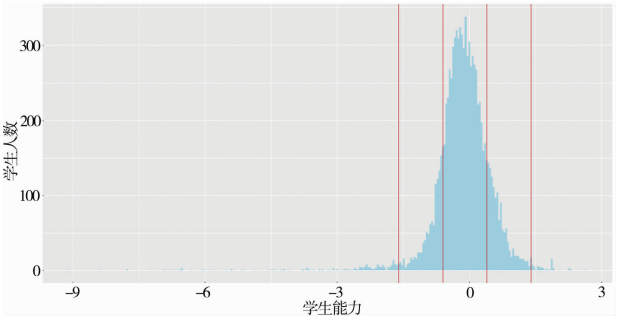


(b) 试题等级分布

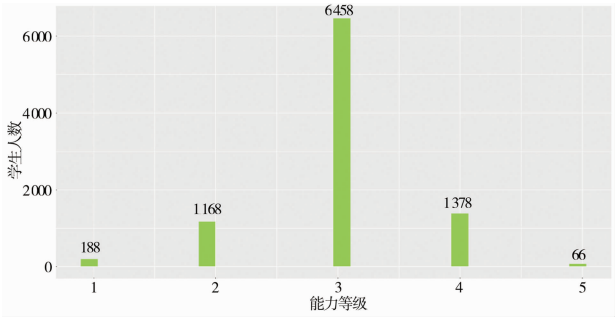
图 3　试题难度及等级分布

图 4(a)是经过本文算法计算出考生的能力值后,通过可视化生成的直方图。由图可见,题目难度以 -2.613 为中心呈正态分布,而能力值以 -0.135 3 为中心呈理想的正态分布。这个结果符合我们的预期。

图 4(b)是将考生数量按能力分成 5 个等级后的柱状图。



(a) 考生能力值分布



（b）考生能力等级分布

图 4　考生能力值及能力等级分布

在对题目难度进行 Rasch 模型迭代计算的同时，本文还计算了题目的 Infit，Outfit 两个参数，这两个参数均用于衡量数学模型和实际数据的吻合度，如公式(9)(10)所示：

$$Var(X_{vi}) = P_{vi} * Q_{vi} \tag{9}$$

$$Z_{vi} = \frac{X_{vi} - E(X_{vi})}{\sqrt{Var(Xvi)}} \tag{10}$$

$$\text{Outfit} = \frac{1}{v}\sum_{v=1}^{V} Z_{vi}^2 \tag{11}$$

$$\text{Infit} = \frac{\sum_{v=1}^{V} Z_{vi}^2 * Var(X_{vi})}{\sum_{v=1}^{V} Var(X_{vi})} \tag{12}$$

式(9)中，X_{vi}为第 v 个考生对第 i 题的作答结果，用 1 表示正确，用 0 表示错误；P 和 Q 分别为答对和答错的概率。式(10)中，$E(X_{vi})$为第 v 个考生对第 i 题作答正确的概率的数学期望；$Var(X_{vi})$为第 v 个考生对第 i 题的作答结果的方差。

对于 Infit 和 Outfit 这两个参数指标，采取以下方式来划分成 3 个等级，其中 1 为最低等级，3 为最高等级。

表 1　Infit，Outfit 等级转换表

数值范围	等级
(2, ∞)	1
(0, 0.8)&(1.2, 2)	2
(0.8, 1.2)	3

本文还计算了题目的区分度(Discrimination Index)，该参数用于衡量一道题目是否能区分优秀的考生和表现较差的考生。将考生根据总成绩由高到低排列，针对每一道题，从作答了这道题的考生中选出总成绩最高的三分之一和总成绩最低的三分之一。具体计算方法如式(13)所示：

$$D_i = \frac{N_{ui} - N_{li}}{1/3} \tag{13}$$

式(13)中，N_{ui}为得分最高的三分之一考生中答对了这道题的人数；N_{li}为得分最低的三分之一考生中答对了这道题的人数。

对于区分度，采取以下方式来划分成 5 个等级，其中 1 为最低级，5 为最高级。

表 2　区分度等级转换表

数值范围	等级
(−1, −0.1)	1
(−0.1, 0.1)	2
(0.1, 0.25)	3
(0.25, 0.4)	4
(0.4, 1)	5

根据 Infit，Outfit 和 Discrimination Index 这三个参数指标所对应的等级，用本文算法能计算出每一道题目的最终鉴别度 J。我们把鉴别度 J 划分成 5 个等级，不仅与考生能力、题目难度相对应，同时方便后续将算法扩展到模拟机的考核成绩分析。

图 5 是经过本文算法计算出试题的最终鉴别度后通过可视化生成的直方图。可以看到题目的鉴别度主要集中在 2 级和 3 级，这部分加起来共有 4 247 道试题，占所有试题的 75%。可见大部分题目的鉴别度并不高，并不能区分成绩优和成绩差的考生。这与题库整体难度较低有一定关联。

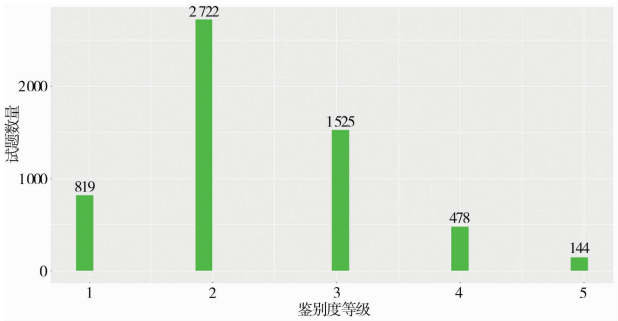


图 5　题目鉴别度等级分布

6　结论

如何根据一个考生的考试结果客观准确地评判其真实水平及飞行能力是目前各大航空公司考试测评系统面临的一个难题。本文首次提出胜任力三层

理论概念,并基于 Rasch 模型计算出较为准确的考生能力、试题难度以及试题鉴别度,进而自动判断试卷可靠性,为考试系统建立了一套自适应的测评方法。目前东航飞培中心通过对飞行员的飞行理论考试提供的真实数据进行的模拟分析结果表明,本文提出的方法能准确反映客观现实,为飞行员能力评价及试卷可靠性评价提供了可行的实施方案以及重要的理论依据。

参考文献:

[1] LORD F M. A theory of test scores[M]. Psychometric Monographs, 1952.

[2] LORD F M. The relation of test score to the trait underlying the test[J]. Educational and Psychological Measurement, 1953,13(4): 517-549.

[3] RASCH G. Probabilistic models for some intelligence and attainment tests[M]. Copenhagen, Denmark: Danish Institute for Educational Research, 1960.

[4] SAMEJIMA F. Estimation of latent ability using a response pattern of graded scores[J]. Psychometrika, 1969,34: 1-97.

[5] HAMBLETON R K, SWAMINATHAN H. Item response theory: principles and applications[M]. Boston, MA: Kluwer Academic Publishers, 1985.

[6] FOX J P. Multilevel IRT modeling in practice with the package mlirt[J]. Journal of Statistical Software, 2007, 20(5): 1-16.

[7] RECKASE M D. Multidimensional item response theory[M]. New York: Springer-Verlag, 2009.

[8] RIJMEN F, TUERLINCKX F, DE BOECK P, et al. A nonlinear mixed model framework for item response theory[J]. Psychological Methods, 2003, 8(2): 185-205.

[9] JABARAYILOV R, EMONS W, SIJTSMA K. Comparison of classical test theory and item response theory in individual change assessment[J]. Applied Psychological Measurement, 2016, 40(8): 559-572.

[10] FISCHER G H, MOLENAAR I W. Rasch models[M]. New York: Springer-Verlag, 1995.

[11] GIN B, SIM N, SKRONDAL A, et al. A dyadic IRT model[J]. Psychometrika, 2020, 85: 815-836.

[12] MARQUARDT K L, PEMSTEIN D. IRT models for expert-coded panel data[J]. Political Analysis, 2018, 26(4): 431-456.

[13] CAMILLI G, GEIS E. Stochastic approximation EM for large—scale exploratory IRT factor analysis[J]. Statistics in Medicine, 2019, 38(21): 3997-4012.

[14] MAIR P, HATZINGER R. Extended Rasch modeling: the eRm package for the application of IRT models in R[J]. Journal of Statistical Software, 2007, 20 (9).

[15] SILVA R M, GUAN Y, SWARTZ T B. Bayesian treatment of non-standard problem in test analysis[J]. METRON, 2019, 77: 227-238.

[16] SILVA R M, GUAN Y, SWARTZ T B. Bayesian diagnostics for test design and analysis[J]. Journal on Efficiency and Responsibility in Education and Science, 2017, 10(2): 44-50.

作者简介

王　真　大学本科,工程师。主要研究方向:飞行理论教学。E-mail: 2316606698@qq.com

Flight theory evaluation method based on Rasch model

WANG Zhen *

(China Eastern Technology Application Research and Development Center Co. ,Ltd)

Abstract: It is very important for airlines and passengers to train qualified pilots. In order to meet the requirements of aviation laws and regulations and flight safety, airlines are required to train and evaluate all pilots regularly. It is of great significance for airlines and training centers to study how to better evaluate pilots through tests, and which indicators can more truly and objectively reflect the level of pilots: flight skills of pilots or their evaluation scores. Based on the Rasch model of IRT theory, this paper establishes a set of adaptive theoretical evaluation methods by analyzing and deriving important parameters such as the difficulty and quality of test questions, the discrimination index of test questions, and the pilot ability, etc. , and then automatically judges the reliability of an exam. At the same time, the regression analysis was carried out with real airline data, so as to establish a set of unique self-adaptive evaluation management methods. And the important concept of Three-Tier of Core Competency was put forward for the first time in the world, which provides a feasible implementation a feasible and valuable theoretical basis for pilot skill evaluation and test reliability assessment.

Keywords: IRT model; Rasch model; item response theory; test analysis module package; maximum likelihood function

* Corresponding author. E-mail: 2316606698@qq.com