基于 IRT 的三年级数学成就测验的编制

范晓玲[1]，魏勇[1,2]
(1.湖南师范大学教科院心理系，湖南 长沙 410081；2.湖南省教育考试院，湖南 长沙 410001)

【摘要】 目的：尝试在项目反应理论的指导下，编制三年级数学成就测验，为学科学习的评价提供一辅助工具。方法：用贝佳方法检验试测数据单维性，利用 ANOTE 软件估计项目参数，采用铆测验设计进行参数等值。结果：四套试卷均符合单维性要求，基本拟合三参数逻辑斯蒂克模型，项目拟合度都在 90%左右，最终选取 52 个项目组成正式测验。结论：正式测验的信息量达到 31.1548，估计标准误为 0.1267，符合项目反应理论的要求。

【关键词】 数学成就测验；3PL 模型；项目参数等值；测验信息量

中图分类号：G449.5 文献标识码：A 文章编号：1005-3611(2006)01-0011-02

Construction of Mathematics Achievement Test for Grade 3rd in IRT

WEI Yong, FAN Xiao-ling
Hunan Normal University, Changsha 410081, China

【Abstract】 Objective: Under the guidance of the item response theory, the author designed the mathematics achievement test and provided a subsidiary test for the evaluation of the performance of mathematics of grade 3. Methods: A two-way specification table was designed before the experiment, The Bejar method was carried out on the data collected, the parameter was evaluated by ANOTE software, the research adopted Anchor-test design, and finished the item parameter equating. Results: All the results were in accordance with the request of unidimensionality, the data fitted Three-Parameter Logistic Model, and finally 52 items were selected to constitute the formal test. Conclusion: The formal test information reaches 31.1548, SE=0.1267, which is accorded with the requirements of the item response theory.

【Key words】 Mathematics achievement test; Three-parameter logistic model; Item parameter equating; Test information

数学是人们生活、劳动和学习必不可少的工具，能够帮助人们处理数据、进行计算、推理和证明；数学在提高人的推理能力、抽象能力、想象力和创造力等方面有着独特的作用[1]。数学成就测验主要用来考查学生学习数学的成就，是根据教材的具体内容和教学目标的要求来编制的，它用于检查学生对数学领域知识、概念、法则等的掌握程度，以及运用这些知识解决实际问题的计算能力。为了帮助教师了解学生是否达到预定的教学要求，及时诊断，及时补救，以实现基础教育的教学目标。本研究借鉴欧美和台湾省基础教育考试的经验，尝试在项目反应理论的指导下，以全日制义务教育数学课程标准为依据编制数学成就测验，旨在为小学三年级数学学科学习评价提供一辅助量表。

1 测验的编制和实施

1.1 测验项目编制

本测验在实地调查及工作分析的基础上，对 1988 年至 2003 年小学三年级数学教材进行系统比较，将教材内容、教育目标及教师经验结合起来，邀请小学数学教师(有三轮以上的教学经验)依据双向细目表，以教材内容为参照物，以教学目标为标尺，以学生实际情况为基础，编写了大量的项目。项目编写完成后，将项目按内容分类编排，形成三年级 ABCD 四套试卷，每套各有项目 80 个。全部项目均采用客观题形式(单项选择题)，以 0、1 记分。

1.2 施测对象

本次测验的被试分别来自长沙市、兰州市、邵阳地区的几所小学，所选学校使用的都是人教版小学三年级数学教材，要求三年级学生全部参加测试。每套测验的试测人数都在三百人以上。因研究时间跨度较大原先测试的三年级学生都已升入四年级，为方便起见，我们仍以三年级称呼。试测时每套测验要求学生在 80 分钟内完成，从实际测试情况来看，学生所用平均时间为 60 分钟左右。

1.3 统计分析工具

本研究利用江西师范大学心理与教育测量中心开发研制的“现代心理与教育测量通用分析系统”(ANOTE1.03) 来估计测验的项目参数与信息函数。用 SPSS11.5 完成其他统计分析。

2 结 果

【基金项目】 省教育厅科学研究项目；感谢杜有志先生的经济资助，感谢参与测试学校的领导与教师。

2.1　模型选择

到目前为止，项目反应理论中已经提出了许多模型，但其中最成熟、最常用的还是单参数、双参数和三参数逻辑斯蒂克模型(One- Parameter, Tow- Parameter, Three- Parameter Logistic Models)。

1978 年，Popham 等人经过实证研究发现，数学成就测验能更好的拟合三参数逻辑斯蒂克模型，使测量误差更小[2]。1989 年，Hambleton 对 1977 年 NAEP 数学成就测验的数据进行参数估计，发现三参数模型能很好的拟合测验数据[3]。

根据前人的研究，结合本测验的实际情况：项目的难度、区分度不同，以及存在被试的猜测。本研究中，我们采用三参数逻辑斯蒂克模型。

2.2　单维性假设的检验

单维性检验的方法比较多，研究中采用影响较大的贝佳方法(Bejar Method)，首先把全部测验项目放在一起进行项目参数估计，然后再从这些项目中选出一部分项目再作一次项目参数估计，比较这部分项目参数的两次估计值，就可直观得了解该测验的单维性[4]。

根据计算，同一批项目两种情况下估计出来的项目参数 b 值的相关系数为 A 卷 0.861**、B 卷 0.841**、C 卷 0.888**、D 卷 0.863**，这是符合贝佳检验要求的，也表明本测验基本上符合单维性要求。

2.3　参数估计

研究中，利用 ANOTE 软件进行项目参数估计和模型－资料拟合检验。参数估计结果显示 A 卷拟合度 90%；B 卷拟合度 90%；C 卷拟合度 92.5%；D 卷拟合度 92.5%。

2.4　项目初选

按照项目反应理论的要求，进行项目初选，首先删除每套试卷中不拟合的项目。另外考虑到本测验项目为客观题，存在猜测可能性，保留值不大于 0.25 的项目，而 b 值与 θ值处在同一取值范围[-3, +3]，a 值越大，区分度越好，但结合前人研究的经验[5]，这里 a 值的取值范围为[0- 2.5]。最后共删除项目 70 个。

2.5　不同测验项目参数的等值

为了将三年级的四个测验转换到同一量表系统上，使得标刻在不同被试群体上的参数有相同的参数量表，进一步完成测验项目的筛选，组拼一套完整的高质量测验。这里需要完成项目参数等值分析，要进行测验等值研究就牵涉到等值数据的采集方法，即等值设计问题。在我们的研究中，A、B 两套测验在同一被试样本组上进行参数估计，由于两测验形式参数在同一组被试上进行估计，其参数本身就标刻在同一量表系统上了。而 C、D 两套施测样本与 A、B 两套施测样本不同，为将 C、D 两套与 A、B 两套转换到同一量表系统上，我们采用了非等组设计——铆测验。

通过 ANOTE 软件我们分别计算了 A 卷与 C 卷、A 卷与 D 卷的变换参数 1=1.0877、1=- 0.6683、2= 1.0236、2=- 0.4812，并完成项目参数的等值转换。

2.6　利用测验信息函数对测验进行评价

经典测验理论用测验信度作为评价测验优劣的一个客观标准，但是经典测验理论信度系数的求取，跟项目难度与区分度是没有关系的，这就给项目选取、测验编制等实际操作问题，留下了许多技术难题；而项目反应理论中信息函数概念的提出，为解决这些困难开辟了一条道路。若测验信息量大，则意味着据此测验考察相应的对象所做推论更可靠，估计误差小。在心理和教育测量中，正是要从被试对项目的做答反应，估计出被试的潜在特质水平，因此若项目性能优良，以此对被试潜在特质水平做出的估计就可靠，误差就会小，项目提供的信息量就大。

这里依据信息函数对测验项目进行再次筛选，挑选出符合要求的项目，最终组成正式测验，根据漆树青等人的研究[6]，确定本测验能容忍的测验标准差应小于等于 0.20，即测验的总信息量应到达 25 以上。

结合测验双向细目表，最后选取 52 个题目组成正式测验。进而计算正式测验的信息量为 31.1548，估计标准误 0.1267，符合项目反应理论对测验质量的要求。结果见图 3-1。

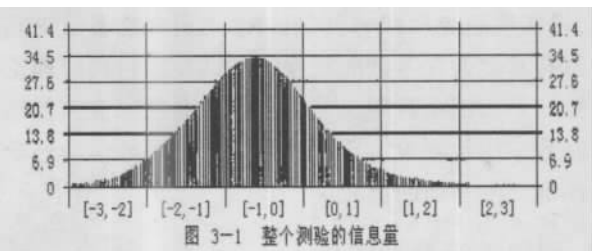

图 3-1　整个测验的信息量

3　结　　论

本研究基本符合单维性假设；项目筛选符合项目反应理论要求；测验信息函数达到项目反应理论的要求，测验信度良好。

(下转第 16 页)

0.8。见表 4。

3 讨　论

本研究表明,16PF 第 5 版的心理测量属性在重测信度和结构效度上较好，而在内在一致性系数方面仍存在问题，尤其是某些因素量表的内在一致性系数未能达到 0.6 的最低标准。根据国内外文献对已有版本 16PF 的研究结果表明,16 个基本因素的内在一致性系数普遍偏低[8]。卡特尔本人对此的解释是量表过高的内在一致性系数可能表明建构量表的题目存在冗余现象，并且他认为非常高的内在一致性系数是不利于测题效度的，因为要达到很高的内在一致性题目必须很相似，从而造成其只能测量范围狭窄而高度具体的某一变量[5]。而如果题目与其所在量表存在相关而与量表内其他题目相关不高的话，表明题目有较高的跨情境的转移性(transferability across situations)，也就是说题目能以一个比较广的范围来对某一变量进行测量。然而大多数理论研究者仍认为测量同一维度题目间高的内在一致性系数是高效度的前提，如果题目间相关过低表明无法确定真正测量的变量，他们认为高信度是高效度的前提，况且 16PF 英文第五版的改进目的之一就是因为人们对 16PF 先前版本低信度的批评,使英文版内在一致性的平均信度达到 0.76,整体范围从 0.68 到 0.76(样本数为 10261 人)。而本研究结果显示,16 个基本因素的平均 α系数为 0.63,整体范围从 0.39 到 0.86,因此就本研究而言,16PF 第五版提高内在一致性信度的目的并未完全达到,提高分量表的内在一致性系数应该成为后续研究的目标之一。

从被试在 185 道题目的反应偏向来看，所有题目都没有超过 90%的判断标准,而且在与被试的面谈中他们普遍觉得题目的跨文化适应性比较好。根据被试样本差异人口统计学变量的检验结果，从性别上看,女性在人格因素上表现得比男性更敏感,这一点与 16PF 第五版技术手册的结论以及先前众多研究的结论是相一致的。根据 16PF 英文技术手册的常模表结果显示,男女在 16 个因素上总共有三个因素存在差异,分别为因素 A,I 和 O,本研究结果之所以只在一个因素上表现出性别的差异可以解释为样本人数偏小以及被试取样群体的特殊性所造成的。按被试来源进行的统计分析结果同样也比较符合实际情况，公司样本中的个体多数已有很多年的工作经验，而且其中相当一部分人是公司的中高层领导,与大学生相比他们的心理成熟度相对较高,情绪稳定,大胆敢为;而多年的工作经历以及作为领导的身份又使他们在规范性上又高于在校大学生。公司工作中员工与员工间，领导与下属间的沟通非常重要，因此乐于沟通而又外向的个性对公司样本中的个体来说很重要，由此造成了公司样本与学生样本间在乐群因素上的差异。可见测题对于不同来源被试的差异检验效果较好，能够比较客观完整地反映不同群体被试间的差异。

参　考　文　献

1 Steven R. Conn & Mark L.Rieke. 16PF Fifth Edition Technical Manual (2nd ed). Champaign, IL: Institute for Personality and Ability Testing, Inc, 1998

2 Mary Russell and Darcie Karol. 16PF Fifth Edition with updated Norms Administrator's Manual (3rd ed). Champaign, IL: Institute for Personality and ability Testing, Inc, 2002

3 金瑜. 心理测量. 上海:华东师范大学出版社, 2001

4 曹小平,任百利,赵泉英,等. 卡氏 16PF 中译本常模 20 余年的变化趋向. 心理科学,1994,173:184-186

5 David JM, Leslie JF. The Psychometric Properties of the 16PF Among Male Anglican Clergy. Pastoral Psychology, 2000, 48(3): 231-240

6 戴忠恒,祝蓓里. Cattell-16PF 修订卡氏十六种人格因素量表手册. 上海:华东师范大学, 1988

7 Cohen J. Statistical Power Analysis for the Behavioral Sciences(2nd ed). Hillsdale, NJ:Lawrence Erlbaum Associates, 1988

8 David JM, Leslie JF. A Comparison of the Psychometric Properties of the 16PF4 and 16PF5 Among Male Anglican Clergy. Pastoral Psychology, 2002, 50(4): 281-289

(收稿日期:2005-06-28)

(上接第 12 页)

参　考　文　献

1 中华人民共和国教育部.全日制义务教育数学课程标准(实验稿).北京师范大学出版社, 2002,1:4

2 Hambleton RK. Application of Item Response Models to Criterion-Referenced Assessment. Applied Psychological Measurement, 1983,7(1): 34

3 Hambleton RK, Hariharan Swaminathan. Item Response Theory: Principles and Applications. Kluwer·Nijhoff Publishing, 1990.172

4 杨志明. 项目反应理论一维性假设的初步研究. 硕士学位论文. 湖南师范大学教育科学学院. 湖南师范大学, 1989. 22-23

5 张敏强, 刘晓瑜.项目反应模型的应用问题研究.心理学报, 1998, 30(4):438

6 漆树青, 戴海崎,丁树良,编著.现代教育与心理测量学原理.南昌:江西教育出版社, 1998.226

(收稿日期:2005-06-20)