不拟合被试对测验信、效度的影响

刘拓，戴晓阳
(深圳大学心理学系，广东 深圳 518060)

【摘要】 目的：考察样本中的不拟合被试对 CTT 框架下测验信度、效度和 IRT 框架下测验信息量的影响。方法：使用 l_z 指标和数据净化的方法，比较不拟合被试删除前后的分半信度、α 系数、测验信息量的变化以及验证性因素分析的效果。结果：①随着删除的不拟合被试比率的增加，E 量表与 N 量表的分半信度、α 系数提高；②删除不拟合被试后，E、N 量表对两因素验证性因素分析模型拟合更好；③不拟合被试的删除，可以提高测验信息量，降低测量标准误。结论：样本中不拟合拟被试的存在，会造成对测验信度系数的低估，影响测验的结构。从 IRT 的角度而言，会造成测验信息量的降低。

【关键词】 项目反应理论；个人拟合；信度；测验信息量

中图分类号：R395.1 文献标识码：A 文章编号：1005-3611(2011)06-0743-03

Effect of Misfit Person on Test Reliability and Validity

LIU Tuo, DAI Xiao-yang
Department of Psychology, Shenzhen University, Shenzhen 518060, China

【Abstract】 **Objective:** To investigate the influence of the inclusion of person misfit on reliability, validity and test information. **Methods:** l_z index and data purification were used to compare the differences in reliability, structure validity and test information between inclusion and exclusion of misfitting person. **Results:** ①As the proportion of deletion of misfitting person increased, split-half reliability and α coefficient improved; ②The removal of misfitting cases data showed that there were better fit indices for the hypothesized structure in confirmatory factor analysis; ③The exclusion of misfitting cases was found to enhance test information and increase the standard error. **Conclusion:** The person-misfit may result in underestimation of reliability, and influence of the structure of test, and may cause the reduction of test information.

【Key words】 IRT; Person-fit; Reliability; Test information

在使用心理学量表进行研究的过程中，研究者往往会根据理论假设，建构出符合实际数据的测量学模型，如 IRT 的 Rasch 模型、等级反应模型(GRM)、部分记分模型(PCM)等。根据这些测量学模型，研究者一方面可以对量表的质量进行分析，另一方面也可以对被试不同的心理特质水平进行评估和解释。通常情况下，数据中大部分被试的反应能拟合某种测量模型，但也有小部分被试的数据会出现无法拟合该测量模型的情况。这些与测量模型不拟合的被试，称为个人不拟合(person-misfit)。研究者认为造成个人不拟合的原因很多，如能力测验中的猜测、创造性作答、不清醒、能力缺失[1]，人格测验中的随机作答、伪装等。

个人拟合指标(person fit index)是探测被试反应模式与测量模型拟合程度的指标。1944 年，Guttman 提出了测量的完美尺度(perfect scale)，成为个人拟合指标的重要理论基础[2]。早期的研究者曾使用相关系数来探测那些不拟合的被试[3]，随着 IRT 理论的发展，越来越多的基于 IRT 模型的个人拟合指标被提出，如 l_z 指标、M 指标、ECI 统计量等[4]。对于诸多个人拟合指标的探测效果，很多研究者进行了比较。如 Rudner 曾比较了 r_{pbis}、r_{bis}、NCI_i、C_i 等指标，认为 U 指标在长测验中效果较好，r_{pbis}、r_{bis} 在短测验中效果好[5]。Karabatsos 比较了 36 种个人拟合指标，认为 H^T 指标对异常反应模式有较好的探测率[6]。个人拟合指标的探测效果也受到多因素的影响，如记分方式、测验长度、项目参数特点等[7,8]。

也一些研究者开始关注个人不拟合被试的作答对测验效果的影响。Schmitt 等人的研究发现在认知测验中，被试的 l_z 值越高效标关联效度就越低[9]。Schmitt 和 Meijer 等人在其系列研究中发现，虽然删除不拟合被试并不能提高效标关联效度，但不拟合被试与效标测验之间呈现出低相关[10,11]。Curtis 的研究认为删除不拟合被试对验证性因素分析效果的影响时好时坏[12]。上述情况表明，关于个人不拟合对测验的影响，研究者尚未得出一致的结论。

国内对于个人拟合的实证研究还较少，曹亦薇曾用 MSD(多维标度法)对不拟合的被试在词汇测验中所选的干扰项特点进行过分类，进而探索这些被试在辨析词义时的思维方式[13]。刘拓等使用 l_z 指标

通讯作者：戴晓阳

探测了人格测验中由于动机缺乏而随机作答的被试，并认为这些被试可能会影响到测验的难度和区分度[14]。

本研究以艾森克人格问卷(EPQ)的结果为例，使用个人拟合指标和数据净化的方法对个人不拟合被试进行筛查。主要目的是：①在经典测量理论的框架下，考察不拟合被试的存在对测验信度和结构的影响；②在 IRT 的框架下，考察不拟合被试的存在对测验信息量的影响。

1 方　法

1.1 数据与模型

本研究使用龚耀先修订的 EPQ 问卷成人版，被试为深圳市 16~18 岁的高中生共 1860 人，其中男生 972 人，女生 888 人。用探索性因素分析对数据的各个分量表进行单维性检验发现，内外向(E)量表与神经质(N)量表可以满足 IRT 分析的条件。参考前人的研究结果与人格测验的特点[15,16]，决定选用 IRT 二参数模型(2PL)作为分析基础。

1.2 个人拟合与数据净化

个人拟合指标有很多，本研究选用了使用较多、效果较好的个人拟合指标 l_z[17,18]。l_z 指标是将 Levine 和 Rubin 提出的 l_0 指标进行标准化后得到的，因此 l_z 渐进标准正态分布，当显著性水平 α=0.05 时，l_z<−1.65 则被判定为不拟合[4]。因为 l_z 指标的计算会受到不拟合被试的影响，可以使用数据净化的方法对数据进行反复的探测删除和放回估计，当参数估计值和 l_z 指标探测率不再发生变化时，则认为不拟合被试基本被剔除，数据净化过程完成。本研究中所有的结果和指标值都是在净化过程完成后得到的。

1.3 使用软件

研究中信度系数的计算使用 PASW18.0，验证性因素分析使用 SAS9.1 中的 PROC CALIS 过程，IRT 分析、l_z 指标的计算和数据净化过程在 R2.12.1 中实现。

2 结　果

2.1 不拟合被试对信度的影响

以 l_z 的值为标准，l_z 值越小说明拟合程度越差。具体分三种情况，删除最不拟合的 50 人、100 人和 150 人，分别考察 α 系数和分半信度的变化情况。为了排除被试人数变化带来的影响，在做分析时还随机删除 50 名、100 名和 150 名被试作为对照组。表 1 对三种情况下的 α 系数和分半信度进行对比。

表 1 结果显示，当删除的被试为不拟合被试时，随着删除人数由 50 人逐步提高到 150 人，分半信度与 α 系数都在逐渐提高；而随机删除相同人数的对照组信度值基本在原始值附近波动，甚至略有下降。说明信度系数的提高是因为删除不拟合被试造成，而不受被试数量变化的影响。删除后的信度系数提高幅度不太大，可能与该样本不拟合人数比率偏小有关(2.69%~8.06%)，但变化趋势是很清晰的。

表 1　删除不拟合被试人数后信度变化的情况

		全数据	删除 50 人		删除 100 人		删除 150 人	
			删除不拟合	随机删除	删除不拟合	随机删除	删除不拟合	随机删除
E 量表	*a* 系数	0.810	0.816	0.811	0.822	0.808	0.827	0.809
	分半信度	0.814	0.819	0.813	0.826	0.811	0.829	0.812
N 量表	*a* 系数	0.859	0.864	0.858	0.869	0.857	0.873	0.858
	分半信度	0.862	0.865	0.861	0.872	0.861	0.875	0.861

表 2　被试删除与两因素验证性因素分析模型拟合情况

指标	全数据	删除不拟合后	随机删除后
GFI	0.893	0.916	0.887
Chi-Square	4300.970	3183.690	4022.558
RMSEA	0.044	0.038	0.045
AIC	2412.970	1295.690	2134.558

2.2 不拟合被试对测验结构的影响

验证性因素分析是考察测验结构效度的方法。本文首先使用 E 量表与 N 量表的 l_z 指标探测结果，删除被判定为不拟合的 226 名被试。然后，随机抽取相同数量的被试，进行删除处理。分别使用全部数据、删除不拟合被试后的数据和随机删除被试后的数据拟合二因素模型，结果如表 2。

表 2 结果显示，模型拟合指标和模型比较指标都显示，删除不拟合被试后模型拟合效果更好，如按 GFI 大于 0.9 的标准，在删除不拟合被试后 GFI 达到 0.916，随机删除被试后甚至有所下降为 0.887。

2.3 不拟合被试对测验信息量的影响

在 IRT 框架中，用信息量概念取代了传统意义上的信度概念。在经典测量理论中，信度反映的是误差对测验(或分测验)的影响；但在 IRT 中，信息量代表的是误差对每名被试在每道题上的影响，将各道题信息量累加就是测验的信息量，测验信息量平方根的倒数就是测验的标准误。因此，对于不同心理特

质水平的被试,测验的评估准确性是不同的。图 1、图 2 是删除不拟合被试前后两分测验信息量的变化情况图。

从图 1 与图 2 可以看出,删除不拟合被试后,两个分测验的信息量都提高了,E 量表最大提高了 1,而 N 量表最大提高了 2。从特质水平估计的准确性看,内外向水平在-1.5~0.5 之间,神经质水平在-0.5~1.5 的被试估计将更加准确。

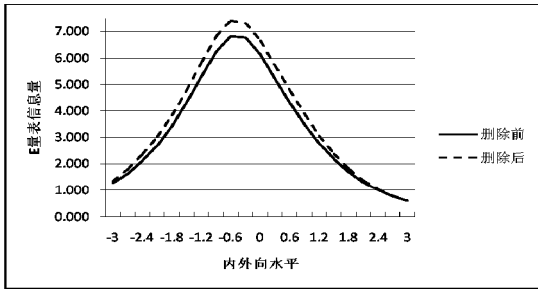

图 1 删除不拟合被试前后 E 量表测验信息函数图

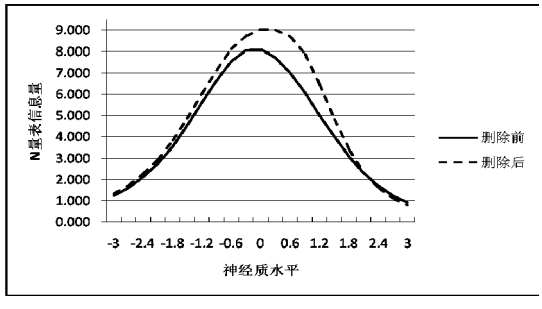

图 2 删除不拟合被试前后 N 量表测验信息函数图

3 讨 论

本研究的结果说明,被试样本中不拟合被试的存在会造成测验信、效度的低估。信度和效度的偏低一方面反映了测验的理论建构不清晰、项目质量不理想;另一方面也可能提示数据的质量存在问题,比如样本中可能含有大量的不拟合被试。因此,在量表编制和修订时,使用个人拟合指标来进行数据的预处理是一个不错的提高数据质量的方法。

另外,从 IRT 的角度看,不拟合被试将会造成测验信息量的降低,导致测量标准误的升高,换而言之,就是会影响到被试评估的准确性。Dodeen 和 Hamzeh 的研究就认为,个人不拟合被试会造成分班或安置性测验中,被试群体的错误分配[19],Hendrawan 等人也认为不拟合被试会造成对被试能力的错误判断[20]。所以删除不拟合被试能提高测验信息量,也可以使得被试心理特质的评估更加准确。

本研究结果只是从变化趋势上证明了不拟合被试对测验信、效度的影响。然而,样本中不拟合被试所占比率的大小对测验信效度影响的程度如何? 以及它们间相互变化的关系等这些问题还有待继续探讨。另外,研究只使用到了 E、N 两个分量表的数据,这虽然与 EPQ 本身的编制方法有关,但结论的推广性还需要更多模拟研究和实证研究的证据。

参 考 文 献

1 Meijer RR. Person fit research:An introduction. Applied Measurement in Education, 1996, 9(1): 3-8

2 Guttman L. A basis for scaling qualitative data. American Sociological Review, 1944, 9(2): 139-150

3 Donlon TF, Fischer FE. An index of an individual's agreement with group-determined item difficulties. Educational and Psychological Measurement, 1968, 28: 105-113

4 Meijer RR, Sijtsma K. Methodology review: Evaluating person fit. Applied Psychological Measurement, 2001, 25(2): 107-135

5 Rudner LM. Individual assessment accuracy. Journal of Educational Measurement, 1983, 20(3): 207-219

6 Karabatsos G. Comparing the aberrant response detection performance of thirty-six person-fit statistics. Applied Measurement in Education, 2003, 16(4): 277-298

7 Reise SP. Scoring method and the detection of person misfit in a personality assessmen context. Applied Psychological Measurement, 1995, 19: 213-229

8 Reise SP, Due AM. The influence of test characteristics on the detection of aberrant response patterns. Applied Psychological Measurement, 1991, 15(3): 217-226

9 Schmitt N, Chan D, Sacco JM, et al. Correlates of person fit and effect of person fit on test validity. Applied Psychological Measurement, 1999, 23: 41-53

10 Schmitt N, Cortina JM, Whitney DJ. Appropriateness fit and criterion-related validity. Applied Psychological Measurement, 1993, 17: 143-150

11 Meijer RR. Person fit and criterion-related validilty: An extension of the schmitt, cortina, and whitney study. Applied Psychological Measurement, 1997, 21(2): 99-113

12 Curtis DD. Person misfit in attitude surveys: Influences, impacts and implications. International Education Journal, 2004, 5(2): 125-143

13 曹亦薇. 异常反应模式的识别和分类. 心理学报,2001,6:558-563

14 刘拓,曹亦薇,戴晓阳. 个人拟合指标在艾森克人格测验中的应用. 中国临床心理学杂志,2011,19(3):323-326

15 Ferrando PJ, Chico E. Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. Educational and Psychological Measurement, 2001, 61(6): 997-1012

(下转第 762 页)

对捐赠意愿的影响完全是通过共情的中介作用来实现的，而且认知和情绪性反应都有效应。该结果与支持 Baston 共情-利他主义理论的研究是一致的，个体不幸情境下同情心的激发是很重要的机制[4,7-9]。这说明当捐赠对象为具体的不幸个体时更能激发捐赠者的同情心。这一点与定险峰等的结果有明显差异，即在群体灾难情境下，人格宜人性对慈善捐赠的影响主要是一种直接的作用，而非共情的中介效应[12]。可能帮助对象为模糊的群体时相对较难激发同情心。另一方面，情境强度（受助者的不幸程度）的影响机制与人格很不一样。情境强度对捐赠意愿影响主要是直接效应，占 71%；共情的间接效应只占小部分，仅为 29%，而且只有认知性反应，同情心并没有中介效应。情境强度最终对捐赠额度的影响更是直接效应占主导地位，占 83%。该结果说明情境强度对慈善捐赠的影响主要不是通过共情为中介实现的，这一点与定险峰等是一致的，情境性因素对慈善捐赠的影响更多的是一种直接效应。人格与情境对慈善捐赠的影响除了机制上不同，作用大小也有较大的差异。相比较而言，人格宜人性和情境强度对捐赠意愿的影响是同样的，但后者对捐赠金额的最终影响占主要地位。

参 考 文 献

1 Croson R, Shang J. The impact of downward social information on contribution decisions. Experimental Economics, 2008, 11: 221-233

2 Martin R, Randal J. How Sunday, price, and social norms influence donation behavior. The Journal of Socio-Economics, 2009, 38: 722-727

3 Eckel C, Grossman P. Subsidizing charitable contributions: A natural field experiment comparing matching and rebate subsidies. Experimental Economics, 2008, 1: 234-252

4 Graziano WG, Habashi MM, Sheese BE, et al. Agreeableness, empathy, and helping: A person × situation perspective. Journal of Personality and Social Psychology, 2007, 93: 583-599

5 Burnstein E, Crandall C, KitayamaS. Some neo-Darwinian decision rules for altruism: Weighing cues for inclusive fitness as a function of the biological importance of the decision. Journal of Personality and Social Psychology, 1994, 67: 773-789

6 Batson CD, Chang J, Orr R, et al. Empathy, attitudes, and action: Can feeling for a member of a stigmatized group motivate one to help the group? Personality and Social Psychology Bulletin, 2002, 28: 1656-1666

7 Twenge JM, Baumeister RF, DeWall CN, et al. Social exclusion decreases prosocial behavior. Journal of Personality and Social Psychology, 2007, 92: 56-66

8 Batson CD. Not all self-interest after all: Economics of empathy-induced altruism. In Cremer D De, Zeelenberg M, Murnighan JK. Social psychology and economics. Mahwah, NJ: Lawrence Erlbaum Associates, 2006. 281-299

9 Batson CD. Empathy-induced altruistic motivation. Draft of lecture/chapter for inaugural Herzliya symposium on "prosocial motives, emotions, and behavior," The 1st Herzlyia symposium on personality and social psychology. Herzliya, Israel, 2008. 24-27

10 Batson CD, Moran T. Empathy-induced altruism in a prisoner's dilemma. European Journal of Social Psychology, 1999, 29: 909-924

11 Batson CD, Ahmad N. Empathy-induced altruism in a prisoner's dilemma II: What if the target of empathy has defected? European Journal of Social Psychology, 2001, 31: 25-36

12 定险峰，易晓明. 群体灾难下的慈善捐赠- 共情的中介效应. 中国临床心理学杂志，2011，19(3)：363-366

13 Davis MH, Soderlund T, Cole J, et al. Cognitions associated with attempts to empathize: How do we imagine the perspective of another? Personality and Social Psychology Bulletin, 2004, 30: 1625-1635

14 De Vignemont F, Singer T. The empathic brain: How, when and why? Trends in Cognitive Sciences, 2006, 10: 435-441

15 陈晶，史占彪，张建新. 共情概念的演变. 中国临床心理学杂志，2007，15(6)：664-667

（收稿日期：2011-05-06）

（上接第 745 页）

16 Reise SP, Waller NG. Fitting the two-parameter model to personality data. Applied Psychological Measurement, 1990, 14(1): 45-58

17 Nering ML, Meijer RR. A comparison of the person response function and the lz person-fit statistic. Applied Psychological Measurement, 1998, 22: 53-69

18 Drasgow F, Levine MV, Mclaughlin ME. Detecting inappropriate test scores with optimal and practical appropriateness indices. Applied Psychological Measurement, 1987, 11: 59-79

19 Dodeen H. The use of person-fit statistics to analyze placement tests. Chicago, 2003

20 Hendrawan I, Glas CAW, Meijer RR. The effect of person misfit on classification decisions. Applied Psychological Measurement, 2005, 29: 26-44

（收稿日期：2011-06-06）