DOI:10.13718/j.cnki.xdsk.2020.05.012

奥尔维斯欺负量表的Rasch模型分析

赵福菓[1],何 壮[1],袁淑莉[1],黄希庭[2]

(1. 贵阳学院 教育科学学院,贵阳 550005;2. 西南大学 心理学与社会发展研究中心,心理学部,重庆 400715)

摘 要:奥尔维斯欺负量表是国内青少年校园霸凌领域应用最广、影响最大的研究工具之一。该工具中文版发布已满20年,在中国被试中的测量属性、该工具在新时代的适用性需要再度检验。本研究应用中文版量表对2 116名学生被试进行施测;使用Rasch模型对数据进行分析。结果表明,量表对国内霸凌行为的调查仍有借鉴意义,但在等级选项设置、题目难度、区分度、项目功能差异、题目内容设计、时代适用性等方面均存在一定程度的问题。奥尔维斯欺负量表是为了调查霸凌现状而编制的工具,将其用于霸凌相关的心理建模研究可能是对该工具的误用。

关键词:OBVQ;Rasch;项目分析;校园霸凌

中图分类号:B841 **文献标识码**:A **文章编号**:1673-9841(2020)05-0115-07

一、引 言

校园霸凌是中小学校学生个体或群体受到力量较强一方蓄意或恶意、持续多次的身心攻击行为。校园霸凌有恃强凌弱、蓄意伤害、重复发生的本质特征,对霸凌者、受害者及旁观者的身心都存在不同程度的影响和危害,扭曲其对教育、社会的看法进而改变正常的行为方式,甚而导致受害者长期抑郁、失去或放弃生命,霸凌者也会遭到报复。对霸凌现状的调查是法律法规、政策制定的重要前提。奥尔维斯欺负量表(Olweus Bully/Victim Questionnaire, OBVQ)是国际公认发布最早、最权威的调查工具之一,迄今为止已经在多个国家、多种语言文化背景下得以应用。

OBVQ由瑞典人Olweus于20世纪80年代编制,中文版由张文新和武建芬于1999年修订并引入中国[1],是国内霸凌相关研究常用的工具之一。仅2017—2019近3年期间,发表在CSSCI和中文核心期刊上的相关研究中,就有14篇应用。相关研究可以划分为三类:第一类是作为调查工具,进行现状调查,如狄文婧等人基于该量表调查了青海省小学生校园霸凌的现状,并对藏汉两地进行了对比[2];第二类是作为霸凌者/被霸凌者/置身事外者的鉴别工具,如桑青松等人将被霸凌维度得分前25%的被试确定为高受霸凌组[3],凌辉等人基于霸凌/被霸凌维度得分区分卷入者身份[4];第三类是将OBVQ作为重要变量,建立中介、调节模型,以解释霸凌及相关变量间的关系,如赵占峰等人研究了青少年同伴侵害与问题行为的关系及心理素质在其中的中介和调节作用[5]。

OBVQ中文版引进已满20周年,我国的社会、经济、文化都发生了巨大变化。校园霸凌的主体、类型、频率等特点是否都与当年有了较大差别,量表是否还适用于今天的被试群体都是需要解答的问题。

收稿日期:2020-02-22

作者简介:赵福菓,贵阳学院教育科学学院,教授。

基金项目:国家社会科学基金教育学一般项目"青少年校园霸凌的特点和机制研究:以贵州省为例"(BBA170070),项目负责人:赵福菓。

二、研究方法

(一)研究工具

本研究以张文新和武建芬于1999年修订的中文版为工具,共分为被他人霸凌(Victim)及霸凌他人(Bully)两个维度,分别包括6个李克特5级量表形式的题目。

(二)样本

2019年10—12月,在贵州省贵阳市的初中、高中、职业高中分层整群抽样,共发放问卷2 177份,收回有效问卷2 116份,有效率97.2%,平均年龄16.4±1.3岁,年龄跨度12～18岁,男女比例4∶6。

(三)数据分析与管理

数据分析基于Rasch理论,Rasch模型是一簇模型的统称,被广泛应用在考试与问卷数据分析中。本研究根据数据的点,选择了Rasch理论一系列模型中的评定等级量表模型(Andrich Rating Scale Model, RSM),该模型由David Andrich于1978年提出,专门用于分析等级量表数据。RSM除继承了Rasch模型参数不变性、参数估计不受被试能力分布影响、精确估计每个项目的测量误差等优点以外,还将数据分析拓展到李克特量表等级设置科学性的评价上。本次数据分析采用的软件为Winsteps 3.74,被霸凌和霸凌他人两维度分别进行。

三、数据分析结果

(一)单维性检验

单维性是指测量过程中有且仅有一种心理特质在影响被试作答。具体到本研究是指学生仅基于霸凌/被霸凌行为的情况作答,答题过程中未受到社会称许效应、主试及其他因素的影响。

Rasch模型通过对测量残差的主成分分析来判断数据的单维性。根据Raiche的建议,首对比残差的特征值应当在[1.4～2.1]之间[6];方差数据中能被Rasch模型解释的比例越高越好。同时Linacre的建议根据被试态度和题目难度来确定方差数据中应被Rasch模型所解释的比例[7]。霸凌维度首对比残差为1.9,由Rasch模型解释的残差比例为28.5%,比Linacre建议的10%～20%之间更高。被霸凌维度首对比残差为1.4,由Rasch模型解释的残差比例为30%,达到了Linacre建议的20%～30%之间的标准。

(二)等级选项分析

1. 被霸凌维度

作答选项按照发生频次或频率由低到高依次是“从来没有过”“总共一两次”“一个月两三次”“一周一次”“一周几次”。相邻两个选项之间的时间间隔不是等距递增的。例如“每周一次”代表平均每7天发生一次,“每月两三次”代表平均每十天发生一次,“总共一两次”则未明确具体时间间隔。若等级选项的设计与被试群体霸凌行为的实际频率不匹配,被试可能会因为发生的频率更高或更低而无法给出准确信息,导致量表不能对被试进行有效区分。例如,被试每天都遭受语言侮辱,则应设置相应选项“每天多次”,否则他只能在当前选项中选择“一周几次”,随后的数据分析,他将被判定为一周受到几次语言霸凌,最终结论数据反映的发生频率远低于实际情况。

使用Rasch模型可以系统地分析每个选项的测量特性。绘制选项概率曲线(Category Probability Curve, CPC)可以判断是否存在选项等级的滥用或缺失。以被霸凌维度为例,图1所示:图中每条曲线对应一个选项,横轴代表被试受到霸凌的程度(从左往右递增),纵轴代表被试选择的概率。由于Rasch模型分析过程中对数据做了中心化处理,因此横轴量尺以0为中心向正负两端无限延伸,数值大小仅代表受霸凌程度高低。

以某位受霸凌程度为−2的被试为例,他选择“从来没有过”的概率约为80%,选择“总共一两次”的概率约为20%,选择其他选项的概率接近于0。据此推断,该被试选择“从来没有过”的可能最大。以此类推,受霸凌程度在$-\infty$,−0.6区间内,即A点左侧的被试,选择“从来没有过”的概率

最大；－0.6，0.2 区间内，即 AE 点之间的被试，选择“总共一两次”的概率最大；0.2，＋∞区间内，即 E 点右侧的被试，选择“一周几次”的概率最大。无论在哪一区间，“一个月两三次”“一周一次”被选择的概率都非常低，两条曲线均被“一周几次”曲线覆盖。结合表 1 数据，测量过程中，有多个等级选项使用率偏低。

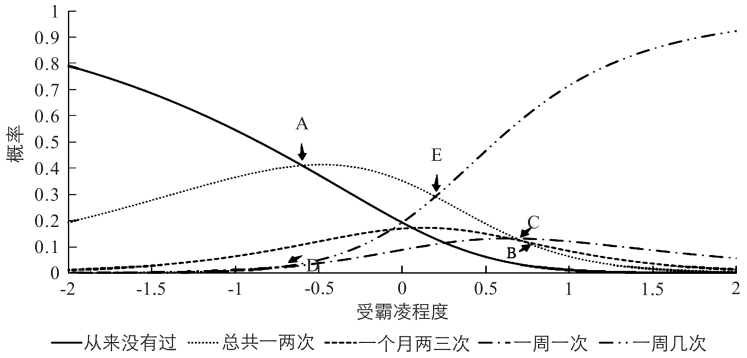

图 1　被霸凌维度 CPC 曲线(五级量表)

Rasch 模型将相邻两个选项曲线的交点称为阈值(Threshold)。以 A 点为例，对应到横轴为－0.6，受霸凌程度低于－0.6 的被试(即 A 点左侧)选择“从来没有过”的概率最高，在 A 点右侧，即受霸凌程度高于－0.6 的被试选择“总共一两次”的概率最高。李克特量表等级代表的含义是递增的，与之对应，阈值也应当是依次递增的。分析结果如表 1，四个阈值排序为 D＜A＜C＜B，顺序颠倒，与李克特量表的基本假设不符。

模型预测与实际数据的一致性也是评价等级选项设置合理性的重要指标，如表 1 所示。M 代表由模型根据受霸凌程度预测出的被试作答情况，C 代表被试的实际作答。表 1“M→C”列表示“预测会出现在某一选项里的作答，在实际测量中仍出现在该选项的百分比”[8]。两个变量的一致性比例越高，则数据与模型假设契合度越高，说明量表的等级设定越合理。“一个月两三次”“一周一次”两个选项的 M→C 比例都在 20％以下；说明等级设置不合理，需要修订。

表 1　被霸凌维度参数估计(五级量表)

选项	阈值	M→C
从来没有过		82％
总共一两次	－0.60(A)	39％
一个月两三次	0.72(B)	15％
一周一次	0.64(C)	15％
一周几次	－0.77(D)	75％

根据 Linacre 的建议，当出现阈值顺序颠倒、李克特等级滥用等情况，应当将相应的选项与相邻选项合并[9]。将“一个月两三次”“一周一次”“一周几次”合并为“多次”，合并后的 CPC 曲线如图 2 所示，代表三个选项的曲线均存在一个区间，在这个区间内，该选项被选择的概率最大，说明测量过程中每个选项都发挥了区分作用。

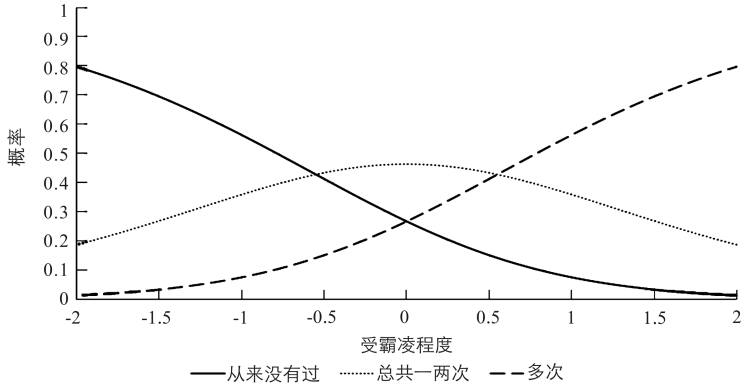

图 2　被霸凌维度的选项概率曲线(三级量表)

修订后三个等级的 M→C 一致性比例有较大程度提高。说明被霸凌维度更适合使用李克特 3

等级量表。五级量表拟合较差的原因可能是被试群体对时间频率的判断能力较差，经常受到他人霸凌的被试很难准确得回忆并报告霸凌事件发生的频率。

表 2　被霸凌维度参数估计(三级量表)

选项	阈值	M→C
从来没有过		82%
总共一两次	−0.55	43%
多次	0.55	68%

2. 霸凌维度

霸凌维度的 CPC 曲线如图 3 所示，从图形上看，代表“一个月两三次”“一周一次”两个选项的曲线均被其他曲线覆盖，未起到区分不同程度霸凌者的作用。

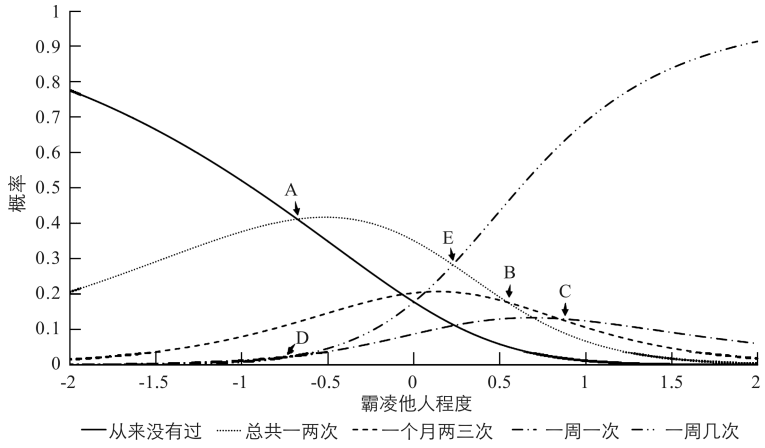

图 3　霸凌维度的 CPC 曲线(五级量表)

霸凌维度阈值出现了顺序颠倒的情况，排序为 D＜A＜B＜C，如表 3 所示。“一周一次”“一个月两三次”“一周几次”选项的 M→C 一致性比例较低。

表 3　霸凌维度参数估计(五级量表)

选项	阈值	M→C
从来没有过		86%
总共一两次	−0.68(A)	45%
一个月两三次	0.54(B)	31%
一周一次	0.84(C)	6%
一周几次	−0.71(D)	＜1%

将“一个月两三次”“一周一次”“一周几次”合并为“多次”，修订后的 CPC 曲线如图 4 所示。

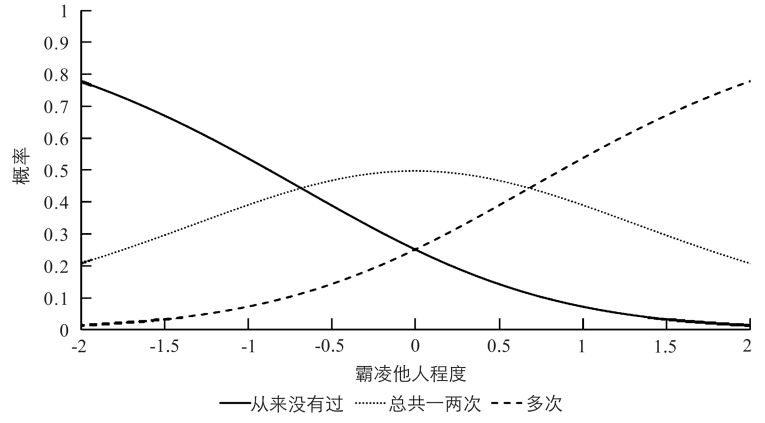

图 4　霸凌维度的 CPC 曲线(三级量表)

合并后的三个等级的 M→C 一致性比例较修订前有较大程度提高，见表 4。综合相关分析结果，认为霸凌维度更适合使用李克特 3 级量表。

表 4　霸凌维度参数估计(三级量表)

选项	阈值	M→C
从来没有过		86%
总共一两次	−0.68	52%
多次	0.68	42%

(三)题目拟合

题目拟合指数通过比较实际数据与模型预期数据的一致性来评价单个题目符合模型假设的程度。常用的拟合指标如表 5 所示，A4、A7、B2、B3、B4、B5、B7 的拟合较差，超出了 Wright 建议的[0.8,1.2]范围[10]。霸凌维度 Infit MNSQ 均值 1.3±0.32；被霸凌维度 Infit MNSQ 均值 1.13±0.24。相比而言，被霸凌维度的拟合均值更接近理想值 1，且标准差更小，拟合更好。霸凌维度拟合均值已超出了建议的范围，说明整个维度的拟合较差。

表 5　题目拟合及 DIF 检验

维度	题号	题目拟合		DIF M−H 检验		
		难度	Infit MNSQ	DIF 值(女−男)	χ^2	p
被霸凌	A2	−0.685	1.08	0.504	4.277	0.040
	A3	−0.164	1.04	−0.719	4.234	0.040
	A4	1.034	1.48	2.027	1.821	0.177
	A5	−0.189	0.86	1.444	6.969	0.008
	A6	−0.313	0.91	0.698	1.571	0.210
	A7	0.317	1.41	2.434	0.831	0.362
霸凌	B2	−1.380	1.29	0.546	6.255	0.012
	B3	−0.612	1.50	−0.890	2.836	0.092
	B4	0.369	1.76	1.318	1.766	0.184
	B5	0.277	0.77	0.480	3.947	0.047
	B6	1.024	1.04	−0.515	2.193	0.139
	B7	0.322	1.45	0.067	0.621	0.431

(四)测量信度

Rasch 模型通过被试分隔系数(Person Separation Index，PSI)、分隔信度(Person Separation Reliability，PSR)和分隔指数(Strata)评价测量信度。分隔指数 Strata=(4 * PSI+1)/3，例如当 PSI=2 时，Strata=3，即量表可以将被试区分为高分、中分、低分三组；同时，PSR 应高于 0.8[10]。实测数据两个维度的 PSI、PSR、Strata 均低于 0.1。说明量表区分度很差，出现这种情况的原因可能是：题目的数量少，且难度(题目中行为的霸凌/被霸凌强度)跨度较小，但被试霸凌/被霸凌程度差异较大，二者间的匹配程度较差。

Rasch 模型还可以通过 Targeting 值来评价题目难度与被试能力的匹配程度；Targeting=题目的平均难度−被试平均能力，该值越小代表二者的匹配越好。两个维度的 Targeting 分别为 3.28 和 4.129，说明题目的平均难度远高于被试群体的平均霸凌/被霸凌程度。

以被霸凌维度难度值最大的 A4 题(我曾被打、踢、推、恶意对待，或者锁在室内)为例，难度为 1.034，96.6%的被试选择“从来没有过”，2.6%选择“总共一两次”，另外三个选项仅 0.8%。该题将几种伤害程度和发生概率不同的霸凌行为合并是导致其难度较大的原因。这表明 OBVQ 部分题目所涉及的霸凌/被霸凌行为代表性不高，在被试群体中发生的概率较低，若想提高量表区分能力，应该再增加一些出现频率较高的行为，或将几种典型行为分别提问。

(五)项目功能差异

Rasch 模型下，在控制了被试特质水平后，比较不同组别间作答概率的差异，若存在显著差异，则认为该题目存在项目功能差异(Differential Item Functioning，DIF)。根据 Zwick 等人的建议，采用 Mantel-Haenszel 法检验性别 DIF，当性别差异的绝对值大于 0.5 且 $p<0.05$ 时认为题目存在 DIF[11]。检验结果如表 5，A2、A5、B2、B3、B5 存在性别 DIF。两个维度上均有较大比例题目存在性

别功能差异，影响了测量的公平性。

四、讨　论

（一）“霸凌”概念内涵的跨文化差异

OBVQ 在中国被试中应用出现的问题可能与“霸凌”的东西方文化差异有关。与西方不同，中文词源学角度提供的证据表明，“重复发生”不是霸凌的界定性特征[12]。但 OBVQ 的指导语中明确“只有重复发生的行为才能称为霸凌”。其次，中文与 Bully 对应或意思相近的词汇有“欺负”“欺凌”“霸凌”“欺辱”“凌辱”等，这些词汇在强度甚至内容上均存在较大差异，在引进过程中未做相应的考虑。这些因素都影响 OBVQ 跨文化的适用性。

等级选项分析显示量表存在等级选项过多且前后含义不统一的情况，前两个等级询问事件发生的次数、后三个等级询问事件发生的频率。这类量表的回答过程需要被试回溯式时距估计能力、自传体记忆的共同参与，多数被试时距估计相对不准确[13]。CPC 曲线和阈值分析发现学生很难在三个代表频率的选项上做出精确判断。这可能与学生的主观感受有关，令他们印象深刻的首先是被霸凌时的心理感受，而非霸凌行为发生的精确时间和频率。因此，进一步假设，仅调查行为发生的次数可能会对数据拟合有所改善。于是，我们将三个代表频率的选项合并为“多次发生”，合并后的分析结果也支持了这一假设。

（二）典型霸凌行为

典型霸凌行为的代表性对量表影响极大，研究发现，部分题目的质量较差或许与霸凌行为的选择有关。如测量身体霸凌的题目 A4（我曾被打、踢、推、恶意对待，或者锁在室内）、B4（我曾经撞、踢、推他/她，或者将他/她锁在室内），将“踢、打、推”同“恶意对待”并列；但“恶意对待”并不一定表现为身体霸凌。且在同一个题目中询问多种霸凌行为发生的情况也会因被试作答时无法对不同行为作出明确区分，导致题目拟合较差。

典型霸凌行为的代表性还体现在是否与社会时代背景有密切的联系。Olweus 发布量表的年代，典型的霸凌类型包括身体霸凌、关系霸凌、言语霸凌。但随着社会发展，新的霸凌形式如“网络霸凌”开始出现；某些霸凌形式受到更多关注，如“性霸凌”“种族霸凌”。OBVQ 并未将上述霸凌行为纳入其中，这会导致基于 OBVQ 的相关研究结论完整性、代表性不足，研究结论的科学性受到挑战。

典型行为的选择还关系到测验的公平性。研究发现，有相当一部分题目存在显著的性别差异，同等霸凌/被霸凌程度下，男女生的得分存在显著差异。这意味着这些题目所选择的典型霸凌行为跨性别一致性较差，基于 OBVQ 开展的性别比较研究都将受其影响。

（三）题目难度分布与区分度

将题目难度与霸凌/被霸凌程度放在同一量尺下比较是 Rasch 模型的优点之一。研究发现，OBVQ 两个维度绝大多数题目为霸凌/被霸凌程度较轻的行为，且难度分布非常集中，导致量表对不同霸凌/被霸凌程度被试的区分度较差，尤其在高霸凌/被霸凌群体中的区分能力不足。

（四）对量表修订或编制的启示

核心概念内涵的文化差异是导致 OBVQ 部分测量学指标较差的一个重要原因，如“重复性”等核心概念细节上的差异将会影响到操作性定义，进而影响测量工具的应用效果。

OBVQ 设计之初，以调查霸凌现状为首要目的，希望尽可能收集与霸凌/被霸凌相关的信息，所以将发生频率作为重要的内容之一。但在国内青少年群体中的应用表明，被试对频率的估计能力较差，且这种状况对测量结果的影响很大。因此，在量表修订或编制适用于中国被试测量工具的过程中，可以充分考虑国内青少年面对霸凌行为时的心理特点，合理设置选项等级内容及数量，尝试以“心理感受”为测量指标，如主观感受到被侵犯的程度：“非常严重”“比较严重”“轻微”“无”。

部分题目拟合较差的原因可能与同一题目包含多种典型霸凌行为有关，在修订或编制测量工

具的过程中，可以考虑将这些典型行为分开考察。并增加一些侵犯程度较严重的霸凌行为，拓宽题目的难度跨度，以提高量表的区分能力，如“网络霸凌”“性霸凌”等有关的行为。并且，对于所选行为，都应当进行性别、民族、城乡、学段等变量的项目功能差异检验，以保证工具的跨群体公平性。

OBVQ有三个特点值得借鉴：一是量表不仅调查霸凌行为发生的频率，还进一步询问这些行为发生的细节以及青少年的态度和反应；二是同时测量霸凌行为与被霸凌行为，为研究霸凌行为提供了全面、丰富的信息；三是量表并非单纯的调查工具，而是Olweus校园霸凌预防项目（Olweus Bullying Prevention Program，OBPP）的内容之一，量表调查的目的是为后续的干预工作及效果评估提供直接证据。

（五）对OBVQ使用的思考

在Olweus的研究和著作中，OBVQ仅作为现状调查的工具出现。受限于心理测量学技术的发展和普及，原版和中文版均未在结构方程模型、项目反应理论框架下进行测量学属性的分析，区分度、效度等特点尚不明确。近年来，国内外研究将OBVQ作为构建心理模型的工具，应用到高级统计分析过程中，且未对数据质量进行检验。这种将调查问卷当做标准化量表的使用方式可能是对OBVQ的误用。

参考文献：

[1] 张文新，武建芬．Olweus儿童欺负问卷中文版的修订[J]．心理发展与教育，1999(2)：8-12.

[2] 狄文婧，陈振宁．校园欺凌行为的族际比较研究——以青海省藏汉两地小学生为例[J]．民族教育研究，2018(2)：24-30.

[3] 桑青松，李海澜，刘思义，等．心理韧性集体咨询对校园受欺凌小学生状态焦虑的影响[J]．心理与行为研究，2019(3)：333-339.

[4] 凌辉，李光程，张建人，等．小学生亲子关系与校园欺凌：自立行为的中介作用[J]．中国临床心理学杂志，2018(6)：1178-1181.

[5] 赵占锋，刘广增，李淑芬，等．青少年同伴侵害与问题行为的关系：心理素质的中介和调节效应[J]．西南大学学报(社会科学版)，2018(5)：90-97.

[6] G. Critical eigenvalue sizes in standardized residual principal components analysis[J]. Rasch measurement transaction, 2005(1): 1012.

[7] LINACRE J M. Variance in data explained by Rasch measures[J]. Rasch measurement transactions, 2008(1): 1164.

[8] 莫慕贞，黄英华，姚静静，等．使用Rasch测量模式分析小学生《教学反馈量表》的心理属性[J]．心理学探新，2012(5)：387-396.

[9] LINACRE J M. Optimizing rating scale category effectiveness[J]. Journal of applied measurement, 2002(1): 85-106.

[10] WRIGHT B D, MASTERS G. Rating scale analysis[M]. Chicago: Mesa Press, 1982.

[11] ZWICK R, THAYER D T, LEWIS C. An empirical Bayes approach to Mantel-Haenszel DIF analysis. Journal of educational measurement, 1999(1): 1-28.

[12] 陈光辉．跨文化心理现象的词源学考证：以欺负现象为例[J]．华东师范大学学报(教育科学版)，2014(3)：93-98.

[13] 杨莲莲，黄希庭，岳童，等．回溯式时距估计的计时机制[J]．心理科学进展，2018(8)：1374-1382.

责任编辑　曹　莉

网　址：http://xbbjb.swu.edu.cn