心理学探新 2023,Vol.43,No.6,558-566
PSYCHOLOGICAL EXPLORATION

量表数据中不努力作答的识别和清理*

王 丹 刘红云
(北京师范大学心理学部,应用实验心理北京市重点实验室,心理学国家级实验教学示范中心[北京师范大学],北京 100875)

摘 要:在对不努力作答(IER)概念和方法综述的基础上,通过一个心理量表的真实测试数据,介绍了如何综合采用不同的方法甄别 IER。主要包括:(1)探索不同指标对不努力作答模式的敏感性,探讨了应用多种指标的必要性以及如何选取的问题;(2)分析 IER 对测验工具指标计算结果的消极影响;(3)总结不努力作答数据清洗方法及注意事项,为提升量表数据质量提供了数据清理方面的建议。

关键词:不努力作答;量表数据;不努力作答模式;数据清理
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003-5184(2023)06-0558-09

1 引言

量表调查是指通过制定详密的工具,要求被调查者据此进行回答以收集资料的方法。因其具有成本低、快捷高效和操作便捷等优点,被广泛应用于心理学、教育学和社会学研究。尽管研究者可以通过量表收集到大量有价值的数据,但是并不能保证作答者参与的热情和动机,如大量研究发现现实中作答率呈现逐年下降的趋势(Anseel, Lievens, Schollaert, & Choragwicka, 2010; Christian, Dillman, & Smyth,2008;Weiner & Dalessio,2006)。特别是当下在线问卷的流行,很难保证在无人监管的情况下作答者认真参与调查(Pauszek, Sztybel, & Gibson, 2017)。

不努力作答(Insufficient Effort Response,IER)又称不认真作答,指被调查者缺乏作答动机,作答不专心、疲劳或加速作答,导致作答数据无法反映其真实物质(Curran,2016;Hong,Stee dle,& Cheng,2020, Huang,Liu,& Bowling,2015;Meade & Craig,2012)。不努力作答的程度在 2% 到 50% 之间(Johnson, 2005;Meade & Craig,2012),不努力作答更容易出现在题量较多的量表中,被调查者的疲劳效应会促使其在后半部分不认真作答(Berry et al.,1992;Clark, Gironda,& Young,2003)。被不努力作答污染的数据,不仅会令个体作答数据无效,还会为量表指标的计算带来偏差,得出不可靠的分析结果(Crede, 2010;Johnson,2005;Huang,Curran,Keeney,Poposki, & DeShon,2012;Maniaci & Rogge,2014;McGonagle, Huang,& Walsh,2015;Merritt,2012;Steedle,Hong,& Cheng,2020;Woods,2006)。由此可知,对不努力作答的甄别就显得很有必要。

目前对不努力作答的主要甄别方法和指标大约有十几种,研究者主要集中在针对量表作答过程中可能出现的不努力作答的行为模式,构建不同的指标并评估指标甄别效果(Dunn,Heggestad,Shanock, & Theilgard, 2018; Huang et al., 2012; Meade & Craig,2012)。面对不断出现的指标方法,如何选择和应用效果最好的指标成为了研究的重点。有研究通过模拟不同程度的不努力作答数据,评估在不同条件下,各指标的敏感度和特异性(Hong & Cheng, 2019;Meade & Craig,2012)。但是,不努力作答表现多样,并非全是随机作答数据,有的还会呈现多种的规律,很难通过模拟数据研究得出的结论对实际测验中不努力作答情况进行推断。也有研究者以实际量表数据为分析对象,评估清除不努力作答数据后,工具质量指标的计算结果的变化(Steedle et al., 2019),但是并未对不同甄别指标在实际数据中的选择进行更进一步的研究。

综上,当前研究大多介绍不努力作答甄别的方法及其效果,而对于实际研究中如何综合应用多个指标进行甄别等问题缺少关注。本文在对不努力作答识别指标进行归纳的基础上,重点探索不同指标的适用性和一致性;并探讨了不努力作答的消极影响。最后,通过比较实际应用中清理不努力作答数据的不同方法,在方法选取方面给出了建议。

* **基金项目**:国家自然科学基金项目(32071091)。
通讯作者:刘红云,E-mail:hyliu@bnu.edu.cn。

2　不努力作答识别方法概述

不努力作答的识别方法分为主动侦查法、作答过程指标法和指标分析法三类。主动侦查法是一种在测验实施之前，通过主动设置题目对不努力作答行为进行识别的方法，主要包括陷阱题、直接反应题和自评准确率。第二类是作答过程指标，基于计算机在线测试的普及，被调查者的作答过程信息可以被轻易获取，比如作答时间和作答完成率。第三类指标分析法是对不努力作答的一类事后甄别方法，该类方法通过计算已回收数据的各项指标，判断被调查者不努力作答的可能性，常用的指标有七种。详细见表1。

表1　不努力作答方法介绍

方法	内容	定义
主动侦查法	陷阱题（Catch Questions）	对那些可能性极小的行为进行询问，比如“我连续两周不睡觉”。
	直接反应题（Directed Response Items）	要求作答者直接给出特定的反应，比如“请选择强烈同意”。
作答过程指标	自评认真程度	要求被调查者评价自己整体作答的可靠程度（Meade & Craig，2012）。
	题目缺失率	由于缺乏作答动机，导致题目缺失（Steedle et al.，2019）。
	作答时长	完成量表的时长（Meade and Craig 2012）。
指标分析法	平均绝对差（Mean Absolute Difference，MAD）	指的是正向题和反向题平均分差的绝对值（ACT，2016）。
	心理测量同义词/反义词指标（Psychometric Synonyms/Antonyms，PS / PA）	正/负相关系数绝对值最大的30对题目组，然后计算每位被调查者在30对题目对上的相关系数（Goldberg & Kilkowski，1985）；或选择相关系数低于－0.6的题目对组成题目组（Meade & Craig，2012）。
	连续相同作答长字符串（LongString Index）	指的是连续选择相同答案的题目数量（Johnson，2005）。
	个体作答变异指标（The Individual Response Variability Index，IRV）	作答者的原始分数的标准差（Dunn et al.，2018）。
	马氏距离（Mahalanobis Distance，MD）	在题目数量维度空间下个体作答和平均作答的距离。
	奇偶一致性指标（Even Odd Consistency Index，Even Odd）	奇数题目和偶数题目的相关一致性（Jackson，1976）。
	标准化对数似然个人拟合指标（Standardized Log－Likelihood，lz）	基于项目反应理论的一种个人拟合（person－fit）统计指标，用于计算每个被调查者期望作答和实际作答的差异（Drasgow，Levine，& Williams，1985）。

3　研究问题与设计

主要采用指标分析法，辅助主动侦查法和作答过程指标，探讨不同方法的应用情况。

3.1　研究方法

3.1.1　测验工具

通过一个实际的网络测试的量表数据，幽默风格量表（Humor Styles Questionnaire，HSQ），对不努力作答识别方法和效果进行研究。HSQ是由Martin等人开发的用于测试幽默类型的5点评分量表（Martin，Puhlik－Doris，Larsen，Gray，& Weir，2003），共有4个子量表，每个量表8道题目。

3.1.2　数据

所用的数据来自于心理测量项目公开的资源（https://openpsychometrics. org/_ rawdata/）。1071名被调查者参与作答，其中男性581名，女性477名，缺失13人；年龄范围为14－70岁。在调查最后会询问被调查者作答准确率（Accuracy），即“请对自己作答的准确程度进行0至100的评分”。由于本量表为人格类型的测验，作答准确率和被调查者的能力无关，只和其作答的认真程度有关，因此被调查者汇报的准确率可等同于自评的认真程度。

3.2　指标截断值的确定

在进行指标识别之前，首先需要设置各指标的截断值（Cutoff）。对于主动侦查法和作答过程数据指标，并没有一个明确设置截断值的方法。这里将自评认真程度不高于50%的被判定为不努力作答；题目缺失率（Missing）可考虑采用缺失1道和2道题这两个标准来判定。

对于指标分析法中的多个指标，确定截断值的方法并不同（分析语句见 https://osf. io/wgfhv/）。LongString的截断值采用Johnson（2005）提出的碎石图法，对所有作答者在每个选项上面不同长度的连续作答的频率进行比较，将碎石图的拐点作为截断值，每一个选项对应一个截断值。根据图1，选项2－4的拐点对应的题目数目为4，选项1和选项5的拐点在3或4，因此最终选出四组截断值，分别是（3，4，4，4，3），（3，4，4，4，4），（4，4，4，4，3）和（4，4，4，4，4）。

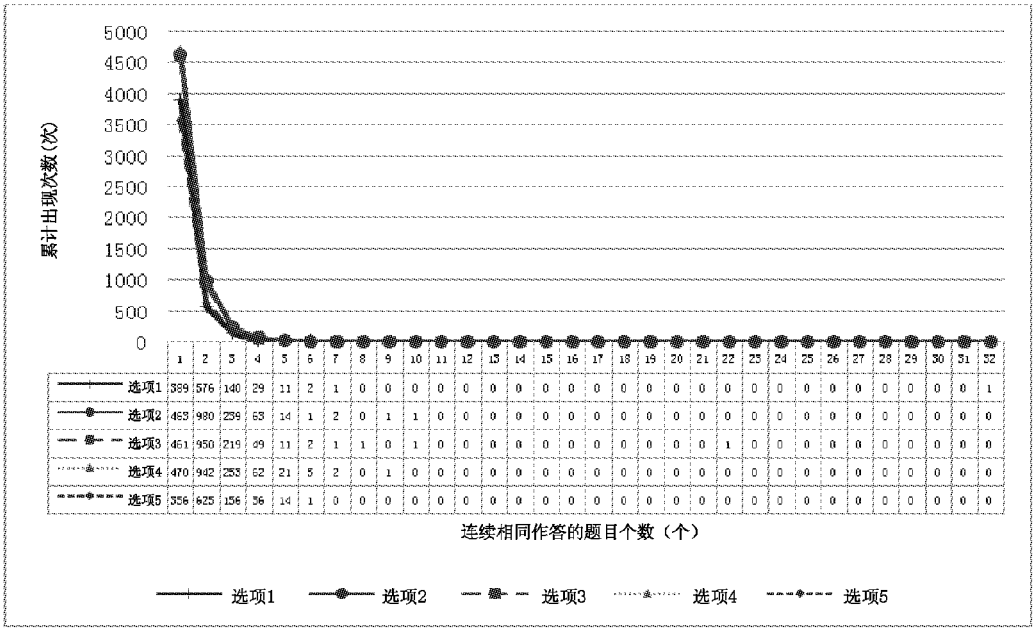

图 1　选项 1 至选项 5 的碎石图

对于 l_z 指标，可直接基于零假设的显著性检验，采用第一类错误率（α）0.01 或 0.05 所对应的临界值作为截断值。利用 R 语言中的 PerFit 包（Tendeiro，Meijer，& Niessen，2016）计算每个被试四个子量表的 l_z 值。参考 Hong 等（2020）的研究，当其中一个子量表的 l_z 低于截断值时，意味着作答者的答案与预期答案的差异在统计上是显著的，代表其很可能没有认真读题或者随机作答，因此被判定为不努力作答。对于 MD 指标，理论上也可以采用零假设的显著性检验，但在实际中，MD 的平方有可能偏离了中心卡方分布，直接采用此方法可能会带来较大偏差（Hong et al.，2020）。

对于 MAD，PS，Even Odd、IRV 和 MD 指标，截断值的确定主要有两种方法。第一种方法是异常值检测，该方法的原理是模拟各指标的零假设分布。首先通过清理数据降低 IER 的消极影响和 α，然后选择合适的 IRT 模型拟合清理后的测验数据，再根据 IRT 参数和能力分布模拟样本作答，根据模拟样本计算出每个指标，并建立该指标零假设下的抽样分布，α = 0.05 和 0.01 对应的值为截断值（Steedle et al.，2019）。第二种方法比较简单，按照比例直接筛选不努力作答，比如 Dunn 等人以 10% 的比例筛选不努力作答被调查者（Dunn et al.，2018），Huang 等人则分别以 1% 和 5% 设置截断值（Huang et al.，2012）。

这里采用第一种方法计算截断值，用 R 语言中的 mirt 包（Chalmers，2012）和 careless 包（Yentes & Wilhelm，2023），语句见附录 1。各项指标的截断值和识别人数见表 2。

表 2　各指标的截断值和识别结果

指标	得到截断值方法	设定 α	截断值	识别人数	占比
MAD	模拟数据	0.01	1.340	43	4.01%
		0.05	1.153	102	9.52%
PS	模拟数据	0.01	-0.155	33	3.08%
		0.05	0.031	133	12.42%
Long String	碎石图		4,4,4,4,4	90	8.40%
			3,4,4,4,4	111	10.36%
			4,4,4,4,3	122	11.40%
			3,4,4,4,3	142	13.26%
IRV	模拟数据	0.01	0.984	120	11.20%
		0.05	1.086	206	19.23%
MD	模拟数据	0.01	53.002	98	9.15%
		0.05	46.872	152	14.19%

续表2

指标	得到截断值方法	设定 α	截断值	识别人数	占比
Even Odd	模拟数据	0.01	-1.000	50	4.67%
		0.05	0.108	102	9.52%
lz	标准正态分布	0.01	-2.330	109	10.08%
		0.05	-1.650	197	18.39%
Missing	缺失数量		≥1	78	7.28%
			≥2	11	1.03%
认真程度	自评		≤50%	16	1.50%

4　不同方法对不努力作答识别的效果及应用

4.1　研究一 IER 指标在不努力作答模式中的适用性

不努力作答的表现形式多样，这里将不努力作答的表现概括为以下五种：

(1)连续相同作答。即连续选择相同答案，比如“3,3,3,3,3,3,3”。

(2)忽略相反题。忽视了当前题目中的相反词，从而出现作答方向错误的情况。

(3)趋中作答。在没有认真阅读题目的情况下，连续选择立场不够明确的中间答案，比如在六点量表中出现大量“3,4,3,3,4,4,4,3,3”模式的作答。

(4)顺序作答。按照顺序选择答案，比如“ABCDABCD……”。

(5)完全随机作答。在不努力作答时，每一个选项都有同等的可能性被不努力作答者选中(Huang et al.,2015)，通常毫无规律。

为了研究不同指标对不同 IER 模式的适用性，针对以上五种不努力作答的模式，就其对应行为的表现特点进行了描述，并在给出了其操作定义见附录表1。

对比每个指标识别出的不努力作答者和努力作答者，在不同模式所对应的操作定义中表现是否有显著差异，从而判断不同指标的模式适用性。采用指标 MAD($\alpha=0.05$)、PS($\alpha=0.05$)、LongString(截断值3,4,4,4,4)、IRV($\alpha=0.01$)、MD($\alpha=0.01$)、Even Odd($\alpha=0.05$)和 *lz*($\alpha=0.01$)区分出的不努力作答群体和努力作答群体在五项行为上的表现，两组群体的平均值和差值在附录的表2中呈现，根据结果可知：

(1)对于连续相同作答，连续相同作答平均长度值越大，说明越容易连续选择相同答案。LongString、IRV 的识别效果较好，识别的出不努力作答者(IER 组)的平均长度值较大，与未识别出的被调查者(安全组)相比差值显著($p=0.025$, *cohen's d* $=0.266$; $p=0.043$, *cohen's d* $=0.223$)。

(2)对于忽略相反题，同一维度下反向题(转换成相同方向后)与正向题分得分方向相反，表明忽略相反题的可能性越大。MAD、PS、MD、Even Odd 和 *lz* 标注出的 IER 组忽略相反题的次数更多，与安全组相比差值都显著($p<0.001$)。根据差值从大到小依次是 MAD、MD、Even Odd、*lz* 和 PS(*cohen's d* 依次为1.589,0.604,0.528,0.547,0.403)。

(3)对于趋中作答，选择“3”的频率越高，说明趋中作答越明显。IRV 指标区分出的两组群体趋中作答的频次差异最大，IER 组与另一组的差值为5.340($p<0.001$, *cohen's d* $=1.493$)，说明 IRV 对趋中作答的识别效果较好。

(4)对于顺序作答，作答数据中顺序作答的数量会较多，说明其按照顺序选择答案的倾向就更明显。IRV 识别效果最佳，IER 组与另一组相比差值为2.140($p<0.001$, *cohen's d* $=0.558$)。

(5)随机作答模式中，以与平均发生率的差值为效标，通常量表中每个选项被选择的频次呈现一定的规律，比如中间选项被选的频次通常较两段的选项高一些，而完全随机作答的数据不会呈现此规律，因此随机选择答案的被调查者的实际选项频率和平均发生率的差值较大。其中差值较大的是 IRV、LongString、MD、*lz* 指标($p<0.001$, *cohen's d* 依次为2.393,1.045,0.265,0.262)。

根据表3可知，IRV 指标比 LongString 表现更好，在一定程度上可以替代 LongString(Dunn et al.,2018)。在“忽略相反题”中，MD 和 *lz* 有不错的表现，因此可与 IRV 组合覆盖全部 IER 模式，达到取长补短的效果。

对不同方法效果之间的一致性进行分析，大部分指标之间的相关系数虽然显著，但是识别效果的并不完全一致。根据表4可知，MD 和 *lz* 之间呈现强相关($r=0.528$, $p<0.001$)，说明二者甄别结果比较一致，二者与 MAD、PS 和 Even Odd 呈现显著正相关；IRV 和 LongString 之间呈现微弱的相关($r=0.161$, $p<0.001$)，二者与其他指标的相关关系并不强，甚至 IRV 与 *lz* 和 MD 呈现微弱的负相关。

表 3 不同 IER 指标的适用情况

指标	类型				
	连续相同作答	忽略相反题	趋中作答	顺序作答	随机作答
MAD		适用			
PS		适用			
LongString	适用				适用
IRV	适用		适用	适用	适用
MD		适用			适用
Even Odd		适用			
lz		适用			适用

表 4 同 IER 指标识别效果的相关系数

	MAD	PS	LongString	IRV	MD	Even Odd	*lz*
MAD	1						
PS	0.129***	1					
LongString	0.046	0.104**	1				
IRV	−0.065*	0.073**	0.161***	1			
MD	0.272***	0.254***	0.147***	−0.092**	1		
Even Odd	0.166***	0.408***	0.109***	0.117***	0.184***	1	
lz	0.272***	0.184***	0.161***	−0.089***	0.528***	0.145***	1

注：*$p<0.05$；**$p<0.01$；***$p<0.001$。

4.2 研究二 不努力作答的识别及其对测验信效度的消极影响

Herman 和 Hilton(2017)认为量表数据质量参差不齐，会对测验工具各项指标的分析产生不可靠的影响。研究二假设删除不努力作答数据之后，会对测验分析提供更准确的工具指标分析结果。在过往的研究中 HSQ 被证明有较好的信效度，是一个稳定有效的测量工具(詹雨臻，陈学志，卓淑玲，& Martin，2011)。对回收的 1071 份数据进行分析，可知 $\alpha=0.862$，CFI = 0.842，RMSEA = 0.060，四个量表的平均相关系数 r = 0.278，与前人研究结果接近。接下来会以 HSQ 的数据为例，演示不努力数据清理的步骤，并比较清理前后的工具指标。首先，进行不努力作答数据的清洗。

第一步，通过主动侦查法，清理不努力作答。本量表没有设置测谎题和陷阱题，只有自我汇报准确率，对于准确率不高于 50% 的数据进行清理；

第二步，通过过程数据清理无效作答。因为缺少作答时间的数据，只能考虑作答缺失，作答缺失在两题及以上的被清理；

第三步，指标识别不努力作答。根据前面研究结果，考虑将 IRV 指标结合 MD 或 *lz* 指标，对不努力作答进行筛选，截断值与前面一致。

值得注意的是前两个步骤的方法对不努力作答的识别虽准确却不够敏感(Meade & Craig，2012)，因此这里将在前两步的基础上结合第三步的指标对不努力作答进行识别，共有六种指标组合，组合 1 没有加入任何指标，是“题目完成率 + 自我汇报准确率”，组合 2 是“题目完成率 + 自我汇报准确率 + IRV”，组合 3 是“题目完成率 + 自我汇报准确率 + MD”，组合 4 是“题目完成率 + 自我汇报准确率 + *lz*”，组合 5 是“题目完成率 + 自我汇报准确率 + IRV + MD”，组合 6 是“题目完成率 + 自我汇报准确率 + IRV + *lz*”。

将原始数据分析得出的工具指标结果作为基线模型，比较六种组合下数据清理后各工具指标的与基线模型的差值。

表 5 数据清理前后的测验工具各项指标平均数(无 IER)

指标组合	IER 比例	内部一致性系数	子量表平均相关系数	CFA 拟合指数	
				CFI	RMSEA
组合 1	2.24%	0.862	0.279	0.843	0.060
组合 2	12.98%	0.865	0.298	0.854	0.057
组合 3	11.30%	0.871	0.280	0.841	0.064
组合 4	12.04%	0.869	0.271	0.866	0.057
组合 5	21.85%	0.875	0.304	0.844	0.063
组合 6	22.50%	0.872	0.292	0.873	0.056

注：基线模型中，Cronbach's $\alpha=0.862$，CFI = 0.842，RMSEA = 0.060，$r=0.278$。

根据表5可知与基线模型相比，各指标组合清洗后的数据所得内部一致性系数和CFI值基本上都更高，大部分组合的RMSEA均小于基线模型。大部分组合的量表平均相关系数也都高于总体。这说明清理了不努力作答数据之后，其描述测验质量相关的各项指标在大部分情况下基本优于不努力作答的数据，工具的信度和效度的指标计算结果变得更好。

不同组合进行比较，组合2和组合6清理后的数据，计算得出的α系数、拟合指数和平均相关系数皆优于基线模型。说明题目作答率、自我汇报准确率、*lz*和IRV在对不努力作答数据清理之后，量表的信度、结构效度和同时效度都能得到更好的验证。

5　讨论和不足

除了介绍不努力作答的方法和类型，以及截断值计算，与以往研究不同的是，对不努力作答的行为模式特点也进行了分类和分析，并在研究一中总结了多种识别指标擅长的不努力作答模式。结果表明IRV属于比较综合的指标，仅在忽略相反题的模式上表现不突出，因此可与在该模式表现较好的MAD、MD、*lz*等指标进行组合筛查。通过各指标识别效果的一致性分析，IRV和MD、*lz*呈现负相关，这可能是因为MD和*lz*主要针对无规律的不努力作答形式，而IRV和LongString则主要针对连续相近或相同作答这类有规律的不努力作答模式。因此，各指标对不同的不努力作答行为各有所长，应当将多个指标综合使用取得最佳甄别效果。研究二演示了不努力作答数据清洗的步骤，结果表明多种方法组合清理后的数据质量更好，将题目完成率、自评认真程度、IRV和*lz*进行组合达到了较好的甄别效果。

不努力作答被认为会对数据分析结果产生消极影响。对比清理前后的作答数据，无不努力作答的数据分析结果显示CFI更高，RMSEA更低，内部一致性系数更好，子量表之间的相关系数也更高。这反映出不努力作答数据对测验工具的信度、结构效度的计算产生消极的影响。努力作答的数据会让分析结果更加稳定，且能更好地拟合量表背后的理论结构，结果也更容易被解释。

对不努力作答甄别方法进行归纳，如表6所示。建议在实际研究中进行不努力数据清洗时，可优先考虑主动侦查法和作答过程指标，因为这些方法是基于被调查者明确的行为，因此更有可靠性，比如作答时间极短的人是无法努力作答的。但这些方法对不努力作答模式不够敏感，检验力有限。比如，由于作答者很容易察觉到预先设置的题目，导致方法失效，因此这类方法识别出的不努力作答者相对其他方法较少（Meade & Criag，2012）；同时作答时间只能找出快速作答者，无法甄别出作答速度正常的不努力作答者。倘若缺乏这类信息或想增加检验力，可考虑使用多种IER指标对作答数据进行事后分析和清洗。

表6　不努力作答数据清洗方法总结与建议

方法	主要指标和方法	优点	缺点	建议
1. 基于测验设计阶段事先设置的题目进行检测	陷阱题 直接反应题 自评认真程度	1. 不容易受到其他因素的干扰，甄别效果较为可靠； 2. 容易计算。	1. 增加作答时长和认知载荷； 2. 有些陷阱题会受到印象管理的影响； 3. 自评认真程度的主观性较强。	1. 结合题目内容，区分不努力和故意作假的情况； 2. 在自评认真程度上采用较严格的标准。
2. 基于作答过程收集的直接信息进行检测	作答时长 作答完成率	1. 计算简便； 2. 无须插入其它题目。	1. 甄别效果不敏感； 2. 有些量表无法提取这类信息； 3. 作答缺失可能并非由不努力导致。	评估题目内容是否会导致作答缺失的情况，进而决定是否能采用完成率来甄别不努力作答。
3. 使用事后甄别指标进行检测	MAD PS/PA IRV LongString MD Even Odd *lz*	1. 无须插入其它题目； 2. 不受量表性质的约束，能够广泛使用； 3. 可根据实际情况灵活选择严格或宽松的标准进行甄别； 4. 每个指标可对应不同的作答模式。	1. 计算过程复杂； 2. 甄别效果容易受到作答倾向的干扰； 3. 没有一个指标能够涵盖所有的不努力作答行为； 4. 阈值的确定没有统一标准。	1. 综合多个指标从不同的角度甄别不努力作答； 2. 建议尝试多种指标，选出最优组合； 3. 需要根据实际情况确定截断值。

根据表6可知，不同方法有各自的优缺点，建议结合多种方法和指标清理不努力作答数据，达到最佳清洗效果。建议采用“MD/*lz* + IRV”指标组合进行甄别，在此基础上也可以再考虑MAD、PS、Even Odd等指标作为补充。

本文主要存在以下两方面的不足。首先，缺乏更加有效的效标对各指标的识别效果进行评估。不努力作答的成因复杂，很难用作答表现直接去解释。

在研究一中,5 种行为仅能说明该被调查者有这样的行为特征,却不能直接说明这样的行为特征完全是由不努力作答引起的,这是存在的局限。其次,根据被调查者自评的准确率可知,不同的被调查者作答认真程度并不相同,目前只是对不努力作答进行了"是"或"否"的区分,却无法评估其不努力作答程度。在后续研究中,对以上两个问题进行深入探讨是有必要的。

6 小结

对不努力作答的常用指标进行梳理,通过一个实际的量表对不努力作答程度以及其消极影响、各指标的具体表现进行了数据分析和探讨,得出以下三个结论:

第一,针对不同的不努力作答行为,不同指标识别效果的并不一致,这反应出不同指标在甄别不同 IER 行为的效果各有所长。

第二,不努力作答会对数据分析结果产生消极影响,不努力作答的数据会导致信度、效度等指标计算结果变差。

第三,针对心理量表的数据,建议综合采用多种方法和多个甄别指标对不努力作答被试进行识别和清理。

参考文献

詹雨臻,陈学志,卓淑玲,Martin.(2011).区分良善与有害的幽默— 正体中文版[幽默风格量表]的发展.*测验学刊*,*58*,207 - 234.

ACT.(2016). *Development and validation of ACT Engage: Technical manual.* Retrieved from https://www.act.org/content/dam/act/unsecured/documents/act - engage - technicalmanual.pdf

Anseel, F., Lievens, F., Schollaert, E., & Choragwicka, B.(2010). Response rates in organizational science, 1995 - 2008: A meta analytic review and guidelines for survey researchers. *Journal of Business and Psychology*,*25*,335 - 349.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K.(1992). MMP 2 random responding indices: Validation using a self - report methodology. *Psychological Assessment*,*4*,340 - 345.

Chalmers, R. P.(2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*,*48*(6),1 - 29.

Christian, L. M., Dillman, D. A., & Smyth, J. D.(2008). The effects of mode and format on answers to scalar questions in telephone and web surveys. *Advances in Telephone Survey Methodology*,250 - 275.

Clark, M. E., Gironda, R. J., & Young, R. W.(2003). Detection of back random responding: Effectiveness of MMPI - 2 and personality assessment inventory validity indices. *Psychological Assessment*,*15*,223 - 234.

Crede, M.(2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*,*70*,596 - 612.

Curran, P. G.(2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*,*66*,4 - 19.

Drasgow, F., Levine, M. V., & Williams, E. A.(1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*,*38*(1),67 - 86.

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N.(2018). Intra - individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*,*33*(1),105 - 121.

Goldberg, L. R., & Kilkowski, J. M.(1985). The prediction of semantic consistency in self - descriptions: characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology*,*48*,82 - 98.

Herman, J., & Hilton, M.(2017). *Supporting students' college success: The role of assessment of interpersonal and interpersonal competencies.* Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/24697

Hong, M., Steedle, J. T., & Cheng, Y.(2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*,*80*(2),312 - 345.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P.(2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*,*27*(1),99 - 114.

Huang, J. L., Liu, M., & Bowling, N. A.(2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*,*100*(3),828 - 845.

Jackson, D. N.(1976). *The appraisal of personal reliability.* Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

Johnson, J. A.(2005). Ascertaining the validity of individual protocols from web - based personality inventories. *Journal of Research in Personality*,*39*,103 - 129.

Maniaci, M. R., & Rogge, R. D.(2014). Caring about carelessness: Participant inattention and its effects on research. *Jour-*

nal of Research in Personality,*48*(1),61 – 83.

Martin,R. A. ,Puhlik – Doris,P. ,Larsen,G. ,Gray,J. ,& Weir,K. (2003). Individual differences in uses of humor and their relation to psychological well – being:Development of the humor styles questionnaire. *Journal of Research in Personality*,*37*,48 – 75.

McGonagle,A. K. ,Huang,J. L. ,& Walsh,B. M. (2015). Insufficient effort survey responding:An under – appreciated problem in work and organisational health psychology research. *Applied Psychology*, *65*, 287 – 321. doi: 10. 1111/apps. 12058.

Meade,A. W. ,& Craig,S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*,*17*(3),437 – 455. https://doi. org/10. 1037/a0028085

Merritt,S. M. (2012). The two – factor solution to Allen and Meyer's (1990) affective commitment scale:Effects of negatively worded items. *Journal of Business and Psychology*,*27*,421 – 436.

Pauszek,J. R. ,Sztybel,P. ,& Gibson,B. S. (2017). Evaluating Amazon's Mechanical Turk for psychological research on the symbolic control of attention. *Behavior Research Methods*,*49*,1969 – 1983.

Steedle,J. T. ,Hong,M. R. ,& Cheng,Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social – emotional learning competencies. *Educational Measurement:Issues and Practice*,*38*,101 – 111.

Tendeiro,J. N. ,Meijer,R. R. ,& Niessen,A. S. M. (2016). PerFit:An R package for person – fit analysis in IRT. *Journal of Statistical Software*,*74*(5),1 – 27.

Weiner,S. P. ,& Dalessio,A. T. (2006). Over surveying:Causes,consequences,and cures. *Getting Action From Organizational Surveys:New Concepts,Methods,and Applications*,294 – 311.

Woods,C. M. (2006). Careless responding to reverse – worded items:Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*,*28*,186 – 191.

Yentes,R. ,& Wilhelm,F. (2023). *Careless:Procedures for computing indices of careless responding.* R package version 1. 2. 2. https://cran. r – project. org/web/packages/careless/index. html

How to Detect and Clean Insufficient – effort Responding in Scale Data

Wang Dan　Liu Hongyun

(Beijing Key Laboratory of Applied Experimental Psychology,National Demonstration Center for Experimental Psychology Education [Beijing Normal University],Faculty of Psychology,Beijing Normal University,Beijing 100875)

Abstract:The current study summarized the methods of insufficient – effort responding(IER) and focused on how to apply these methods to flag IER data in practical research. Taking the data of Humor Styles Questionnaire as an example,two ways of calculating the cutoff of each method are introduced. In the first study,five IER patterns were defined. Comparing the difference between the flag data and Unflag data identified by 7 IER indicator in five patterns,we explored the applicability of each indicators for different IER patterns. In Study 2,we demonstrated the three steps of flagging IER data,and then the negative effects of IER data were investigated by calculating changes in coefficient alpha,correlation coefficients between scales,and confirmatory factor analysis model – data fit. In discussion,we summarized the advantages and shortcomings of every method and gave advices about choosing and using these methods. In conclusion,insufficient – effort responding is prevalent and it can obscure the calculated result of psychometrical indicators. So,it is necessary to identify and clean IER data. This study could help researchers realize the importance of removing IER data and get to know how to identify the IER data in their surveys,mainly including how to calculate the cutoff,and how to choose and apply IER indices.

Key words:insufficient – effort responding;scale data;careless responding;data chean

附录：

表 1 不同 IER 模式所对应的行为表现及操作定义

IER 模式	对应的行为	操作定义
连续相同作答	连续相同作答平均长度	将连续相同作答序列或单独一个作答视为一个单位，32 除以单位数量即为平均长度，最小为 1。
忽略相反题	同一维度反向题与正向题得分相反的次数	将量表中四个子量表的正向题目和反向题目进行比较，方向不一致记为 1，以此类推，四个维度都不同则记为 4。方向一致指的是反向计分后，正向题目的平均分和反向题目的平均分都大于等于 3 或都小于等于 3，则为方向一致。
趋中作答	中间选项出现的频率	统计被调查者选择中间答案“3”的次数。
顺序作答	顺序作答的数量	相连的三个选项为连续数字则算作顺序作答，如 12345、54321，计算顺序作答中数字的数目。
随机作答	与平均发生率的差值	作答者在各个选项的频率与平均频率的差值绝对值的平均数，作答者该选项频率 = 此作答者选择该选项的数目/32，平均发生频率 = 该选项总数/所有选项总数。

注：随机作答模式中，将自评准确率不高于 50% 或存在缺失值的作答数据清理，然后计算剩余作答者各选项的平均频率。选项 1、2、3、4、5 的平均发生频率分别为 0.168、0.224、0.218、0.228 和 0.162。

表 2 各指标对不同类型 IER 行为的识别效果

指标	群体	人数	连续相同作答平均长度	正向题和反向题方向不一致次数	选 3 的次数	顺序作答答案数	随机作答
MAD	unflag	969	1.287	0.570	7.160	4.630	0.083
	flag	102	1.605	1.750	5.400	3.890	0.101
	差值		−0.318	−1.180***	1.760***	0.740*	−0.018***
	95% CI		−0.916 ~ 0.281	−1.323 ~ −1.045	1.040 ~ 2.467	−0.050 ~ 1.535	−0.029 ~ −0.008
Ps	unflag	938	1.309	0.640	6.970	4.580	0.083
	flag	133	1.373	0.970	7.110	4.480	0.100
	差值		−0.064	−0.330***	−0.140	0.100	−0.018***
	95% CI		−0.238 ~ 0.110	−0.465 ~ −0.190	−0.897 ~ 0.614	−0.612 ~ 0.801	−0.026 ~ −0.009
Long String	unflag	960	1.260	0.680	7.070	4.740	0.080
	flag	111	1.808	0.680	6.260	3.050	0.129
	差值		−0.548*	0.000	0.810	1.690***	−0.050***
	95% CI		−1.095 ~ 0.000	−0.176 ~ 0.173	−0.169 ~ 1.793	1.012 ~ 2.377	−0.060 ~ −0.039
IRV	unflag	951	1.268	0.670	6.390	4.320	0.079
	flag	120	1.710	0.780	11.730	6.460	0.132
	差值		−0.443*	−0.110	−5.340***	−2.140***	−0.053***
	95% CI		−0.949 ~ 0.064	−0.257 ~ 0.033	−6.109 ~ −4.578	−2.861 ~ −1.406	−0.060 ~ −0.047
MD	unflag	973	1.282	0.640	7.250	4.730	0.082
	flag	98	1.667	1.160	4.430	2.930	0.110
	差值		−0.385	−0.520***	2.820***	1.800***	−0.028***
	95% CI		−1.008 ~ 0.238	−0.730 ~ −0.326	2.104 ~ 3.533	1.072 ~ 2.528	−0.040 ~ −0.016
Even Odd	unflag	969	1.282	0.640	6.970	4.560	0.083
	flag	102	1.648	1.120	7.180	4.580	0.097
	差值		−0.366	−0.480***	−0.210	−0.020	−0.014*
	95% CI		−0.964 ~ 0.232	−0.695 ~ −0.264	−0.929 ~ 0.514	−0.81 ~ 0.778	−0.024 ~ −0.004
lz	unflag	963	1.281	0.640	7.230	4.740	0.082
	flag	108	1.638	1.110	4.830	3.030	0.109
	差值		−0.356	−0.470***	2.400***	1.710***	−0.028***
	95% CI		−0.921 ~ 0.208	−0.668 ~ −0.283	1.709 ~ 3.085	0.988 ~ 2.429	−0.039 ~ −0.016
total		1071	1.317	0.680	6.990	4.560	0.085

注：差值为不努力作答群体和认真作答群体在该行为表现上的平均分之差。* $p < 0.05$；** $p < 0.01$；*** $p < 0.001$。