**首届全国大学生心理与行为在线实验精英赛**

**（总决赛· 研究报告）**

|  |  |
| --- | --- |
| **研究题目** | **甄别题、反应时与眼动的应用：在线问卷数据清洗新视角** |
| **团队名称** | ***p*<0.05** |
| **参赛赛道** | **A自选赛道 √ B 揭榜赛道 □** |
| **问卷分享链接（Credamo见数平台）** | **作答链接（被试所见）：**  https://www.credamo.com/s/ueu2M3  https://www.credamo.com/s/q6NjY3  **协作链接（供组委会检查，可编辑）：**  详见备注 |

**甄别题、反应时与眼动的应用：在线问卷数据清洗新视角**

**【摘 要】**本研究旨在进一步探讨提高在线问卷信效度的方法。在线问卷作答时，被试更可能不认真或不诚实地回答问题，从而降低问卷的信效度。本研究使用Credamo见数在线平台发布问卷和实验，以内隐联合测验结果为效标，评估在在线问卷中加入甄别题、答题反应时以及眼动对问卷效标效度的影响。研究结果表明，采用认真程度甄别题剔除异常被试数据可能提升问卷效度，在样本量不充足时，采用绝对反应时标准、认真甄别题、测谎题清洗被试数据反而会降低问卷信度。在目前的在线眼动精度不足的现实条件下，采用真实眼动可能降低问卷信度，采用虚假眼动可能降低问卷信效度。本研究为在线问卷数据的质量控制提供了有益的参考和建议。

**【关键词】**在线问卷数据清洗；在线问卷信效度；甄别题；反应时；眼动

# 1 引言

随着互联网的发展，在线问卷发布平台日趋成熟，因此，心理学研究越来越多地采用在线问卷作为数据收集方式（K. B. Wright, 2005）。在线问卷有许多优点，包括但不限于覆盖范围广、省时省力和经济节约。然而，相较于线下一对一收集被试数据，在线收集数据时更难保证被试完成题目时是认真而诚实的。不认真作答与说谎作答的问卷会污染数据，从而大大降低研究结果的真实性（Curran, 2016），尤其是对问卷的信效度可能产生非常负面的影响（DeSimone等, 2018）。

为了探究如何更好地利用在线问卷平台收集问卷，减少被试不认真作答、说谎等情况对研究结果的影响，提高科研效率，本研究将探究利用各类甄别题、答题反应时清洗问卷数据以及在问卷中加入眼动是否能够提升问卷的信度。本研究中的问卷与实验在Credamo见数在线实验平台上发布，因为该平台具有采集眼动数据与记录作答反应时的功能，同时可以在问卷内嵌入简单的行为学实验。鉴于目前大部分在线问卷平台并不具备眼动数据的采集功能，加上采集眼动数据需要在线被试拥有摄像头并允许平台获取摄像头数据等一系列限制，本研究将进一步探究虚假眼动（告知被试将采用其眼动数据，但实际不采集）是否可以提升问卷的信度，从而使研究成果能够推广至不支持眼动数据采集的情况。本研究以在线内隐联想测验结果为效标。

# 2 文献综述

## 2.1 在线收集问卷的优缺点

在线收集问卷对科研人员具有以下优点：首先，它可以让科研人员更容易接触到特定的人群。例如，对于患有特定疾病（如艾滋病）的人群，由于疾病被污名化，患者常常不愿接受线下问卷调查。而在线问卷对于该类患者更具匿名性与隐蔽性，科研人员可以通过这种方式获得他们的真实回答（K. Wright, 2000）。其次，它可以节省科研人员的时间和精力。在线收集问卷时，科研人员可以在短时间内跨越地理限制，接触到数千名具有相同特征的人（Nayak & K A, 2019）。最后，它可以节约研究成本。通过将纸质问卷转为电子形式问卷，可以节省打印成本与邮寄成本。同时，电子问卷可以直接以各类方便分析的格式直接导出，因此也可以节省转录结果的成本（K. B. Wright, 2005）。

然而，相较于线下收集问卷，线上收集问卷也存在以下问题不容忽视：首先，线上收集问卷可能导致抽样问题。例如，在线问卷填写可能存在自我选择偏好，即具有特定特征的被试可能更倾向于填写问卷，这可能导致系统性偏见（Toepoel & Emerson, 2017）。其次，相较于线下问卷，线上问卷具有提交便利（Johnson, 2005）、匿名性强（Meade & Craig, 2012）、作答环境不可控（Barge & Gehlbach, 2012）、主试与被试互动少等特点（Francavilla等, 2019），因此被试更容易不认真作答或做出社会称许性反应，造成污染数据的严重后果。

此外，线上收集问卷还存在其他一些局限性。例如，由于网络环境的不稳定性和技术设备的限制，可能会导致被试在答题过程中出现卡顿、闪退等问题，从而影响数据的准确性和可靠性。同时，线上调查还容易受到样本偏差的影响，因为只有具备一定上网条件的被试才能参与到在线调查中来，这可能导致样本不够随机化，无法完全代表整个研究群体的观点和态度。

因此，在使用线上收集问卷时，研究者需要充分认识到其局限性，并采取相应的措施来降低其对研究结果的影响。例如，可以采用多个题目逐步测试的方法来检查数据的完整性和准确性；同时还可以结合多种数据收集方法来进行综合分析，以提高研究结论的可靠性和可信度。

## 2.2 不认真作答与社会称许性反应

问卷中的不认真作答可能是由于题目难度过高、被试能力不足、被试作答动机不足等原因（Zhong等, 2021），具体表现为被试随机作答（Marjanovic等, 2015）、直线作答（Curran, 2016）、无意义的规律选择答案等模式（Dunn等, 2018），以及多种模式混用。其中，随机作答指被试用掷骰子、随意选择等方式作答，直线作答指被试未经深度思考便选择同一选项（例如：都选择A）的作答方式；无意义的规律选择答案指被试使用与作答无关的无意义规律选择选项（例如：依次选择ABCD并循环）的作答方式；不认真作答对问卷的因子分析结果、信效度等方面会产生严重的负面影响（Kam, 2019）。

社会称许性反应指被试以更为社会所接受的方式作答的倾向，而不是遵循自己真实的想法答题。这可能使结果具有倾向性，威胁测量的效度（Havan & Kohut, 2023）。与不认真作答不同的是，社会称许性反应并不会减少被试的认知负荷（Grau等, 2019）。

传统线下问卷中曾采用设置甄别题的方式以减轻不认真作答与社会称许性反应对问卷造成的负面影响。甄别题包括陷阱题、固定选项题、说谎测试题。其中，陷阱题与固定选项题主要针对不认真作答，说谎测试题主要针对社会称许性反应。陷阱题为正确答案显而易见的题目，例如，“我已经环游世界92次了”（Dunn等, 2018）；固定选项题为要求被试选择特定选项的题目，例如，“本题请选择第二个选项”（Anduiza & Galais, 2017）；说谎测试题，是指那些被试可能因为社会称许性而选择与自身真实想法相违背的选项的题目。例如，“偶尔我会想到一些坏得说不出口的话”。这类题目旨在检测被试在回答问卷时是否如实回答。

近年来，电子问卷兴起，许多研究者也将反应时纳入清洗数据的考虑指标之中。过往研究者常采用绝对标准或相对标准设定反应时阈值。如Soland等人（2019）就以绝对标准“题均2秒”设置阈值。也有研究设定相对标准，例如Mayerl（2013）就采用“平均值±2×SD”设置阈值，Schnell（1994）则采用“Q50 – (1.5×IQR) ”设置阈值，Wise（2017）则通过观察反应时分布图像来确定阈值，当被试群体中有人认真作答，有人不认真作答时，整个群体的反应时会呈双峰分布，当样本足够大，构成双峰分布时，也可以用此方法确定阈值。

## 2.3 内隐联想测验

内隐联想测验（Implicit Association Test，IAT）是一种衡量个体心理内隐态度和联系的工具，它通过测量反应时来揭示人们对不同概念的内隐态度（Greenwald等, 1998）。

IAT的操作方式相对简单，它可以通过测试概念之间的自动关联来测量内隐态度，且效应十分明显（Greenwald等, 1998）。例如，一项针对种族态度的IAT可以揭示人们对不同种族的内隐态度。反应时间短暂的配对可能表明个体对某一种族的正面态度，而较长的反应时间则可能意味着个体存在负面态度（Marini等, 2021）。

近年来，IAT在研究人类态度、偏见和认知之间的关系方面取得了显著进展。研究者们利用IAT来探究种族、性别、年龄等社会身份因素对个体态度的影响。此外，IAT也被广泛用于研究社会问题，如歧视、偏见和隐性偏见等（Colledani & Ciani, 2021; Sohn等, 2021）。通过IAT，我们能够便捷地了解人类的内隐态度和心理联系。

此外，IAT作为一种以反应时为指标的内隐测量方法，具有较高的信度与效度，且能够巧妙地避免自陈报告中存在的社会赞许性、态度防御、自我矫饰等的影响。IAT数据的处理有多种方法，Greenwald发现D值法的效果最好，即先计算不相容任务和相容任务的反应时平均数之差，再除以两个任务上所有反应时的标准差。通过D值法计算的结果不仅具有明显和稳定的内隐效应，而且与外显的主观自陈测量相关。因此，本研究将以IAT的效应量为校标，从而衡量在线问卷的信效度。

## 2.4 眼动与在线眼动测量

眼动追踪可用于测量眼球运动模式及瞳孔大小等眼部特征，从而探测个体观察的位置及其观察行为方式。从前，研究人员需借助眼动仪来记录参与者在观察过程中的眼球运动轨迹。近年来，随着眼动数据收集技术的不断进步，研究者们能够更全面、准确地了解参与者的注意分配和信息处理，网页眼动分析软件的出现也为在线问卷同步收集眼动数据提供了可能（Clark & Stephane, 2018）。在线眼动多为直接使用计算机配备的摄像头进行数据记录，参与者在真实环境中的表现更为自然，有助于增强研究的外部有效性，结果更具有现实意义。

然而，在线眼动数据的质量和可靠性难以保证。因为研究参与者的设备和环境可能各不相同，不同眼动设备、眼动测量方式的算法也各不相同，眼动数据的准确性可能会受到上述因素的影响，未来仍需要更深入的研究与开发从而进一步提升数据准确性。（Zou & Zhang, 2019）。

# 3 问题提出

本研究旨在探讨提升在线问卷信度的方法。在线问卷在实际应用中具有显著的优势，但也存在一些局限性，如难以确保参与者的认真和诚实。为了解决这一问题，本研究提出了两个研究问题：

* 利用甄别题与反应时清洗数据是否能够提升在线问卷信效度？
* 问卷中设置真实眼动是否能够提升在线问卷信效度？问卷中设置虚假眼动（即让被试误以为其作答过程中的眼动数据将被采集且纳入数据分析）是否也能够提升在线问卷信效度？若能，其作用与真实眼动相比是否有显著差别？

本研究分为研究一和研究二两个子研究进行。研究一将回答第一个问题，研究二将回答第二个问题。我们希望通过本研究，能够为提高在线问卷的信效度提供新的视角和建议。

# 4 实证研究

## 4.1 研究一：利用甄别题与反应时清洗数据对在线问卷信度的影响

### 4.1.1 研究目的

1. 探究基于反应时对被试进行筛选是否可以提升在线问卷的信效度。

2. 探究基于甄别题作答情况对被试进行筛选是否可以提升在线问卷的信效度。

### 4.1.2 被试

利用Credamo平台收集698名中国本土长大的、汉语为母语的18 ~ 26岁学历高中及以上的被试，男女各半。

### 4.1.3 研究设计

本实验共分为两个部分：问卷部分与IAT部分。实验中采用IAT实验结果作为问卷效标，分别计算各组问卷得分与IAT结果之间的相关系数。

问卷部分采用2×3双因素混合设计。因素一为是否收集被试的反应时数据，该因素有两个水平，分别为：收集与不收集。因素二为甄别题，该因素有三个水平，分别为：无甄别题、采用认真程度甄别题（包括高/低频题与指定选项题）、采用说谎甄别题（包括态度防御题与测谎题）。因变量为问卷最终得分。被试被随机分为6组。问卷中，所有题目顺序都是随机的，性别刻板印象测量题每题一分，甄别题作答错误的问卷将被整份剔除，存在反应时过短作答（单题小于2.5秒）的问卷将被整份剔除（Huang等, 2012; Soland等, 2019; Zhong等, 2021）。

IAT部分采用单因素被试内设计。自变量为任务性质，共两个水平：相容任务和不相容任务，因变量为反应时。任务顺序在被试间对抗平衡。若被试作答错误，将提示被试作答错误并要求被试即时更正，被试更正的反应时会累计入该试次的反应时。

### 4.1.4 工具及材料

本研究通过Credamo在线问卷收集平台进行，使用材料包括学科性别刻板印象问卷和IAT行为实验。本研究以学科性别刻板印象为主题，是因为相较于其他主题，被试对刻板印象的自陈报告更容易受到社会赞许性的影响，因此被试在作答时更有可能隐瞒真实情况或展现出态度防御。

本研究采用的学科性别刻板印象问卷利用ChatGPT辅助生成。首先，我们利用ChatGPT生成学科性别刻板印象包含的维度。我们对这些维度进行了讨论和筛选，并对部分维度进行了合并。之后，请ChatGPT根据这些维度生成包含30个题项的李克特5点量表。我们仔细斟酌每个题项，对其进行删减和修改。最终，学科性别刻板印象问卷由20个题项构成，包含4个维度，它们分别是：

* **创新和研究维度**，包含5题，主要考察被试是否认为性别会影响学科内的创新和研究方向。例如，“你是否认为某个性别在某些学科领域内更具创新能力？”。
* **领导和代表性维度**，包含4题，主要考察被试是否认为某个学科是由某个性别所领导和代表的。例如，“你是否认为某些学科领域内的领导地位应该主要由某个性别担任？”。
* **性别的成绩差异维度**，包含5题，主要考察被试是否认为某个性别在某个学科中能取得更好的成绩。例如，“你是否感觉不同性别的学生在某些学科的学业表现上存在差异？”。
* **学科选择和倾向维度**，包含6题，主要考察被试是否认为性别会影响某学科的选择。例如，“你是否认为某些学科更适合某个性别？”

IAT实验呈现的刺激材料包含四种词汇：男性词、女性词、理科词、文科词（具体词表见附录），共包含相容和不相容两种情况，共包含7个实验组块，共计180个试次。

### 4.1.5 研究过程

研究过程如图4-1所示。首先，在Credamo在线问卷收集平台上发布问卷，被试被随机分为六组，分别填写对应的三种问卷中的一种。问卷中题目顺序在维度内随机，三种问卷仅在甄别题维度不同：无甄别题组的问卷中无甄别题、认真程度甄别题组的问卷中有3道认真程度甄别题（包括1道高频题、1道低频题、1道固定选项题）、说谎甄别题组的问卷中有3道说谎甄别题。性别学科刻板问卷（不含甄别题）共20题，题目均为李克特五点量表题。被试通过键盘上按键进行作答（按键“1”表示“完全不符合”；按键“2”表示“不太符合”；按键“3”表示“一般”；按键“4”表示“比较符合”；按键“5”表示“完全符合”）。对于需要收集反应时的三组，利用Credamo平台记录被试反应时。

被试填完问卷后，进入IAT实验模块，具体过程如表4-1与图4-2所示。实验中，每组中一半的被试先进行相容任务，后进行不相容任务；另一半被试先进行不相容任务，后进行相容任务。

被试先进行练习，相容任务（或不相容任务）的练习共有两部分组成：（1）概念词（文科与理科）的分类任务；（2）属性词（男性与女性）的分类任务。在进行相容任务练习时，首先进行概念词（文科与理科）的分类任务。此时，屏幕上方一左一右分别呈现一个类别词，左侧为“理科”，右侧为“文科”。而后在屏幕下方中央呈现一个词（如“哲学”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）。概念词分类任务结束后，进入属性词分类任务。同样，屏幕上方一左一右分别呈现一个类别词，左侧为“男性”，右侧为“女性”。而后在屏幕中央呈现一个词（如“妇女”）， 被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）；在进行不相容任务的练习时，首先进行概念词（文科与理科）的分类任务，屏幕上方一左一右分别呈现一个类别词，但与相容任务相反：左侧为“文科”，右侧为“理科”。而后在屏幕下方中央呈现一个词（如“哲学”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）。概念词的分类任务结束后，进入属性词的分类任务。同样，屏幕上方一左一右分别呈现一个类别词，与相容任务一致：左侧为“男性”，右侧为“女性”。而后在屏幕中央呈现一个词（如“妇女”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）。

正式实验则要求被试对概念词和属性词进行联合反应。在相容任务条件下,屏幕上方一左一右分别呈现两个类别词，左侧为“理科或男性”，右侧为“文科或女性”，而后在屏幕下方中央呈现一个词（如“哲学”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）；在不相容任务条件下,同样,屏幕上方一左一右分别呈现两个类别词，左侧为“文科或男性”，右侧为“理科或女性”，而后在屏幕下方中央呈现一个词（如“哲学”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）。正式实验单次试验流程见图4-3。

|  |
| --- |
|  |
| **图4-1　利用甄别题与反应时清洗数据对在线问卷信度的影响研究流程** |
|  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **表4-1　IAT 实验序列** | | | | | | |
| **组别** | **任务性质** | **任务类型** | **试验次数** | **功能** | **左键对应项目** | **右键对应项目** |
| B1 | 相容 | 学科词分类 | 20 | 练习 | 理科 | 文科 |
| B2 |  | 性别词分类 | 20 | 练习 | 男性 | 女性 |
| B3 |  | 联合分类 | 20 | 练习 | 理科或男性 | 文科或女性 |
| B4 |  | 联合分类 | 40 | 正式 | 理科或男性 | 文科或女性 |
| B5 | 不相容 | 学科词分类 | 20 | 练习 | 文科 | 理科 |
| B6 |  | 联合分类 | 20 | 练习 | 文科或男性 | 理科或女性 |
| B7 |  | 联合分类 | 40 | 正式 | 文科或男性 | 理科或女性 |
| **注: 对于一半的被试，B1、B3、B4 组分别与 B5、B6、B7 组互换，即相容与不相容任务顺序互换。**     |  | | --- | |  | | **图4-2　内隐联想测验（IAT）流程序列** | | | | | | | |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
| **图4-3　内隐联想测验（IAT）单次实验流程** |

### 4.1.6 数据处理方法

本研究通过甄别题和反应时行为数据这两种方式剔除了不认真作答或存在态度防御或撒谎倾向的被试。问卷剔除标准如下：（1）剔除离群值（单个样本问卷得分与线性回归预测的问卷得分之差为单个样本的残差。将所有样本的残差进行标准化，标准化后大于3或者小于-3的样本判定为离群）；（2）针对反应时数据，剔除答题时间过短的被试（平均每题答题时长小于2500毫秒）；（3）针对高频甄别题，剔除得分过低的被试（高频题得分小于3分）；（4）针对低频甄别题，剔除得分过高的被试（低频题得分大于3分）；（5）针对指定选项题，剔除未按照指令作答的被试；（6）针对态度防御及说谎甄别题，剔除平均分过低的被试（平均分小于等于3分）；（7）剔除IAT正确率未达75%的被试。

根据以上标准，各组剔除的被试量如下：A1组（无甄别题，无反应时）不删除任何被试， 保留有效问卷115份，有效问卷占该组总问卷的比例为100%；A2组（无甄别题，有反应时）剔除12份问卷，保留有效问卷103份，占比89.57%；B1组（高低频及指定选项题，无反应时）剔除31份甄别题作答错误的问卷，剔除1份离群值问卷，剔除3份IAT正确率小于75%的问卷，保留有效问卷80份，占比69.57%；B2组（高低频及指定选项题，有反应时）剔除31份甄别题作答错误的问卷，剔除1份离群值问卷，剔除3份IAT正确率小于75%的问卷，剔除7份反应时异常的问卷，保留有效问卷73份，占比63.48%；C1组（态度防御及说谎甄别题，无反应时）根据甄别题剔除24份问卷，剔除2份IAT正确率小于75%问卷，保留有效问卷91份，占比77.78%；C2组（态度防御及说谎甄别题，有反应时）根据甄别题剔除24份问卷，剔除2份IAT正确率小于75%问卷，剔除11份反应时异常的问卷，保留有效问卷80份，占比68.38%。

为了衡量问卷的信度，我们计算了问卷的分半信度（Guttman折半系数）以及总问卷和各维度的克隆巴赫系数。为了分析问卷的效度，我们以IAT行为实验中的d值为校标，检测问卷的效标效度。在处理IAT数据时，为了避免个别有问题的试次对整体结果的影响，我们对数据进行如下处理：（1）在正确的试次中，对于反应时低于300ms的试次，反应时转化为300ms；对于反应时高于3000ms的试次，反应时转化为3000ms；（2）在错误的试次中，对于反应时低于300ms的，转化为600ms；反应时高于3000ms的，转化为3300ms；反应时介于300ms与3000ms之间的，在原始反应时基础上增加300ms；（3）剔除反应时大于10000ms的试次；（4）在计算反应时平均数时，错误反应的反应时也计算在内。数据分析采用*d*值法，即先计算不相容任务和相容任务的平均反应时之差，再除以两个任务所有试次的反应时的标准差，从而获得*d*值。在本实验中，由于不区分对于男女偏见的倾向，计算IAT结果和问卷得分相关性时，取了*d*的绝对值|*d*|。|*d*|值越大说明被试对于男女性别和学科的偏见越大

### 4.1.7 数据分析及结果

首先，我们对实验一中收集到的问卷进行了描述性统计，表4-2呈现了实验一中不同条件下问卷的总得分以及各个维度得分的平均值和标准差。在学科性别刻板印象问卷中，各组不论是总得分，还是各个维度的得分均处于3（一般）和4（比较符合）之间，说明总体上被试具有一定的学科性别刻板印象。

我们还对A、B、C三个大组的被试的总得分进行了方差分析，三组被试在问卷上的总得分并没有显著差异（F(2,283) = 0.901, *p* > 0.05, 偏*η*2 = 0.006）。因此，各组被试在学科性别刻板印象上无显著差异，他们都具有轻微的性别刻板印象。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **表4-2****实验一各组别问卷描述性统计** | | | | | | | | | | |
| **组别** | **创新及研究** | | **领导和代表性** | | **性别的成绩差异** | | **学科选择和倾向** | | **总得分** | |
| *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| A1 | 3.23 | 0.67 | 3.25 | 0.76 | 3.29 | 0.92 | 3.55 | 0.72 | 3.34 | 0.55 |
| A2 | 3.17 | 0.65 | 3.17 | 0.74 | 3.24 | 0.89 | 3.49 | 0.72 | 3.28 | 0.52 |
| B1 | 3.37 | 0.49 | 3.20 | 0.58 | 3.18 | 0.88 | 3.64 | 0.47 | 3.38 | 0.42 |
| B2 | 3.38 | 0.51 | 3.20 | 0.58 | 3.22 | 0.86 | 3.64 | 0.48 | 3.39 | 0.42 |
| C1 | 3.45 | 0.66 | 3.25 | 0.65 | 3.32 | 0.80 | 3.61 | 0.52 | 3.43 | 0.46 |
| C2 | 3.42 | 0.66 | 3.25 | 0.63 | 3.35 | 0.77 | 3.62 | 0.53 | 3.43 | 0.46 |

**注: A1为无反应时且无甄别题组，A2为有反应时且无甄别题组，B1为无反应时且有认真程度甄别题，B2为有反应时且有认真程度甄别题，C1为无反应时且有测谎题，C2为有反应时且有测谎题。**

实验一中各个条件下IAT结果*d*值的描述性统计如下所示，表4-3中出现了各个不同条件下*d*值的平均值和标准差，以及样本*d*值平均值小于0的*t*检验值及对应概率*p*。各组中，IAT结果*d*值均显著小于0，说明被试总体上存在学科性别刻板印象，并且总体上被试认为理科更偏向男性，文科更偏向女性。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **表4-3**各条件下d值结果 | | | | | | | |
|  | ***n*** | ***M*** | ***SD*** | ***t*** | ***df*** | ***p*** | ***d*** |
| A1 | 115 | -0.385 | 0.406 | -10.159 | 114 | <0.01 | 0.947 |
| A2 | 103 | -0.375 | 0.417 | -9.133 | 102 | <0.01 | 0.900 |
| B1 | 80 | -0.421 | 0.370 | -10.161 | 79 | <0.01 | 1.136 |
| B2 | 73 | -0.444 | 0.355 | -10.663 | 72 | <0.01 | 1.248 |
| C1 | 92 | -0.400 | 0.363 | -10.576 | 91 | <0.01 | 1.103 |
| C2 | 81 | -0.391 | 0.344 | -10.213 | 80 | <0.01 | 1.135 |

**注: A1为无反应时且无甄别题组，A2为有反应时且无甄别题组，B1为无反应时且有认真程度甄别题，B2为有反应时且有认真程度甄别题，C1为无反应时且有测谎题，C2为有反应时且有测谎题。**

我们还对6组问卷的*d*值进行了双因素方差分析，结果表明甄别题与反应时两个因素的主效应均不显著（甄别题F(2,543) = 0.838，*p* > 0.05；反应时F(1,543) = 0.001, *p* > 0.05），交互作用不显著（F(2,543) = 0.100, *p* > 0.05）。表明本实验的被试总体存在学科性别刻板印象，在三个问卷组中学科性别刻板印象没有显著差异。

我们对实验一各个条件收集到的问卷进行了信度分析，表4-4呈现了各个维度和总得分的克隆巴赫系数以及Guttman折半系数。通过比较，可以发现各组的分半信度均较好（Guttman折半系数均大于0.7），总量表的克隆巴赫系数也较高（均大于0.7）。因此说明总体上问卷的信度较好。然而，细看各个维度的克隆巴赫系数，性别的成绩差异维度的信度较好，但是其他维度的信度较差。比较各个条件，可以发现B、C组的信度在数值上都低于A组的信度，A2、B2、C2的折半系数均略低于其对应的未删反应时组A1、B1、C1，A2、B2的总量表克隆巴赫系数均略低于其对应的未删反应时组A1、B1。这可能是由于B和C组相较于A组，A2、B2、C2相较于A1、B1、C1都剔除了部分问卷，导致样本量减少导致的，具体将在讨论部分进行论述。

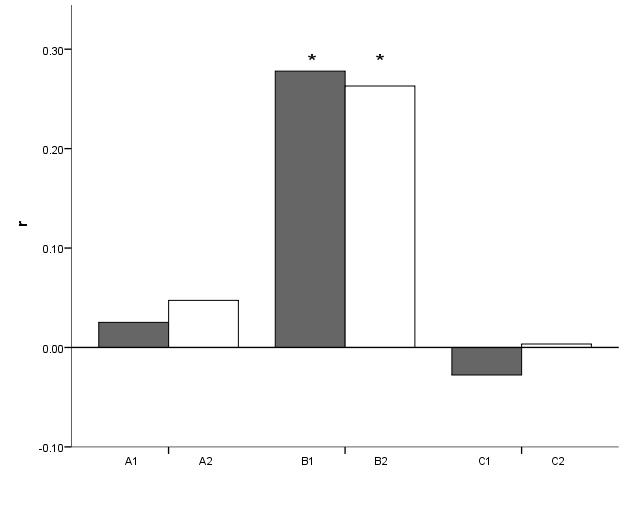
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **表4-4　实验一不同组别问卷信度分析** | | | | | | |
| **组别** | **克隆巴赫** | | | | | **Guttman**  **折半系数** |
| **创新及研究** | **领导和**  **代表性** | **性别的**  **成绩差异** | **学科选择**  **和倾向** | **总量表** |
| A1 | 0.561 | 0.477 | 0.802 | 0.655 | 0.809 | 0.871 |
| A2 | 0.559 | 0.452 | 0.788 | 0.664 | 0.788 | 0.859 |
| B1 | 0.190 | 0.121 | 0.805 | 0.359 | 0.712 | 0.781 |
| B2 | 0.234 | 0.069 | 0.794 | 0.380 | 0.704 | 0.763 |
| C1 | 0.604 | 0.268 | 0.710 | 0.510 | 0.755 | 0.755 |
| C2 | 0.628 | 0.286 | 0.700 | 0.561 | 0.767 | 0.730 |

**注: A1为无反应时且无甄别题组，A2为有反应时且无甄别题组，B1为无反应时且有认真程度甄别题，B2为有反应时且有认真程度甄别题，C1为无反应时且有测谎题，C2为有反应时且有测谎题。**

以IAT行为实验中的|*d*|值为校标，检测问卷的效标效度。实验一中各个条件下|*d*|和问卷得分的相关性如表4-5与图4-4所示。在B1条件下，IAT结果与问卷得分非常显著弱相关；在B2条件下，IAT结果与问卷得分显著弱相关；而在其他组均无显著相关性。说明在B1和B2条件下问卷的甄别效度要好于其他组。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **表4-5　各条件下|*d*|与问卷得分的相关性** | | | | |
|  | ***r*** | ***df*** | ***z*** | ***p*（单尾）** |
| A1 | 0.03 | 114 | 0.270 | >0.05 |
| A2 | 0.05 | 102 | 0.479 | >0.05 |
| B1 | 0.28 | 79 | 2.471 | <0.01 |
| B2 | 0.26 | 72 | 2.232 | <0.05 |
| C1 | -0.03 | 91 | -0.262 | >0.05 |
| C2 | 0.00 | 80 | 0.032 | >0.05 |

**注: A1为无反应时且无甄别题组，A2为有反应时且无甄别题组，B1为无反应时且有认真程度甄别题，B2为有反应时且有认真程度甄别题，C1为无反应时且有测谎题，C2为有反应时且有测谎题。**



**图4-4　各条件下|*d*|与问卷得分的相关性**

**注: A1为无反应时且无甄别题组，A2为有反应时且无甄别题组，B1为无反应时且有认真程度甄别题，B2为有反应时且有认真程度甄别题，C1为无反应时且有测谎题，C2为有反应时且有测谎题。\*代表在α水平为0.05时具有统计学显著性。**

## 4.2 研究二：采用眼动对在线问卷信效度的影响

### 4.2.1 研究目的

1. 探究利用眼动数据进行数据清洗是否可以提升在线问卷的信效度。

2. 探究采用虚假眼动是否可以提升在线问卷的信效度。

### 4.2.2 被试

利用Credamo平台收集168名中国本土长大的、汉语为母语的18 ~ 26岁学历高中及以上的被试，男女各半。

### 4.2.3 研究设计

本实验共分为两个部分：问卷部分与IAT部分。实验中采用IAT实验结果作为问卷效标，分别计算各组问卷得分与IAT结果之间的相关系数。

问卷部分采用单因素被试间设计。自变量为是否采用眼动数据，自变量共三个水平：采用眼动数据、告知被试采用眼动数据但实际不采用、不告知被试采用眼动数据且实际不采用。根据自变量的三个水平，将被试随机分为三组：真实眼动组、虚假眼动组、无眼动组。无眼动组与虚假眼动组的问卷仅开头指导语不同。对于虚假眼动组，被试会被告知他们的眼动数据会被采集，但实际上，我们并不会打开该组被试的摄像头从而记录该组被试的眼动数据。对于眼动组，被试不仅会在指导语中得知其眼动数据会被记录，其还需要进行眼动校准。问卷中，所有题目顺序都是随机的，性别刻板印象测量题每题一分。

IAT部分与研究一相同。

### 4.2.4 工具及材料

被试的作答电脑须配有摄像头。其余材料同研究一。

### 4.2.5 研究过程

整体实验流程如图4-5所示。首先，在Credamo在线问卷收集平台上发布问卷，被试随机分为三组：无眼动组、虚假眼动组、眼动组。被试填写与其分组对应的问卷。问卷共20题，题目均为李克特五点量表题。被试通过键盘上按键进行作答（按键“1”表示“完全不符合”；按键“2”表示“不太符合”；按键“3”表示“一般”；按键“4”表示“比较符合”；按键“5”表示“完全符合”）。

被试填完问卷后，进入IAT实验模块，IAT具体过程基本同研究一，但对于分配到虚假眼动组与真实眼动组的被试，在进入正式实验前，会征求被试“是否同意开启摄像头记录眼动轨迹”。只有同意开启摄像头的被试才能进入后续实验。对于虚假眼动组，我们在指导语中告知被试其注视点轨迹会被记录，但实际并不记录其眼动数据。对于真实眼动组，在其同意开启摄像头后，将进入眼动校准环节，校准率高于等于60%后方可进入正式实验。

|  |
| --- |
|  |
| **图4-5　采用眼动对在线问卷信度的影响研究流程** |

### 4.2.6 数据处理方法

问卷剔除标准如下：（1）剔除离群值；（2）剔除IAT正确率未达75%的被试。A组为无眼动组，未剔除任何被试，保留有效问卷117份，有效问卷占该组总问卷的比例为100%；B组为虚假眼动组，剔除2份离群值问卷，剔除2份IAT正确率小于75%的问卷，保留有效问卷114份，占比96.61%；C组为真实眼动组，剔除1份离群值问卷，剔除1份IAT正确率小于75%的问卷，保留有效问卷110份，占比98.21%。实验中，由于平台对于眼动数据的精度严重不足，我们无法通过眼动数据剔除被试。

IAT实验|*d*|值数据清洗标准与流程同实验1。

### 4.2.6 数据分析及结果

表4-6呈现了实验二的问卷的描述性统计结果。和研究一的结果类似，各组被试在各个维度和总分的得分处于3（一般）和4（比较符合）之间，说明被试具有一定的学科性别刻板印象。方差分析显示三组被试在问卷的总分上并不存在显著差异（F(2,336) = 0.992, *p* > 0.05, 偏*η*2 = 0.006）。

**表4-6　实验二问卷描述性统计**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **组别** | **创新及研究** | | **领导和代表性** | | **性别的成绩差异** | | **学科选择和倾向** | | **总得分** | |
| ***M*** | ***SD*** | ***M*** | ***SD*** | ***M*** | ***SD*** | ***M*** | ***SD*** | ***M*** | ***SD*** |
| A | 3.23 | 0.67 | 3.25 | 0.76 | 3.29 | 0.92 | 3.55 | 0.72 | 3.34 | 0.55 |
| B | 3.28 | 0.69 | 3.15 | 0.67 | 3.31 | 0.83 | 3.55 | 0.63 | 3.34 | 0.51 |
| C | 3.35 | 0.57 | 3.34 | 0.60 | 3.29 | 0.80 | 3.66 | 0.56 | 3.43 | 0.46 |

**注: A为无眼动组，B为虚假眼动组，C为真实眼动组。**

表4-7呈现了实验二中信度分析的结果。通过观察Guttman折半系数以及总量表的克隆巴赫系数可以得知，问卷的分半信度和内部一致性信度都较为不错，Guttman折半系数和克隆巴赫系数均大于0.7。A组的总量表的克隆巴赫系数与Guttman折半系数均大于B组和C组，这可能表明在问卷中增加真实或虚假的眼动测量可能降低问卷信度。C组在除性别的成绩差异维度以外的所有的信度指标上都不高于A组和B组，这可能也与繁琐的眼动矫正过程导致被试的认知资源不足有关，具体将在讨论部分进行论述。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **表4-7　实验二问卷信度分析** | | | | | | |
| **组别** | **克隆巴赫** | | | | | **Guttman**  **折半系数** |
| **创新及研究** | **领导和**  **代表性** | **性别的**  **成绩差异** | **学科选择**  **和倾向** | **总量表** |
| A | 0.561 | 0.477 | 0.802 | 0.655 | 0.809 | 0.871 |
| B | 0.515 | 0.217 | 0.701 | 0.507 | 0.760 | 0.818 |
| C | 0.447 | 0.084 | 0.779 | 0.493 | 0.760 | 0.803 |

**注: A无眼动组，B为虚假眼动组，C为真实眼动组。**

以IAT行为实验中的|*d*|值为校标，检测问卷的效标效度。实验二中A、B、C三个条件中的*d*值结果如表4-8所示。各组中，IAT结果*d*值均显著小于0，说明被试总体上存在学科性别刻板印象，并且被试总体上认为理科更偏向男性，文科更偏向女性。

**表4-8　实验二各条件下d值结果**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ***n*** | ***M*** | ***SD*** | ***t*** | ***p*** | ***d*** |
| A | 115 | -0.385 | 0.406 | -10.159 | <0.01 | 0.947 |
| B | 114 | -0.422 | 0.337 | -13.392 | <0.01 | 1.254 |
| C | 110 | -0.437 | 0.333 | -13.787 | <0.01 | 1.315 |

**注: A无眼动组，B为虚假眼动组，C为真实眼动组。**

我们还对A，B，C三组问卷的*d*值进行了单因素方差分析，结果表明三组之间的*d*值不存在显著差异（F(2,439) = 1.013，*p* > 0.05）。表明本实验的被试总体存在学科性别刻板印象，在三个问卷组中学科性别刻板印象没有显著差异。

实验二中各个条件下|*d*|和问卷得分的相关性如表4-9所示。在B条件下，IAT结果与问卷得分显著弱负相关，而在其他组均无显著相关性。说明在告知被试会监测其眼动，但随后并未跟随有眼动校准时，被试作答的认真与诚实程度会更低。

**表4-9　各条件下|*d*|与问卷得分的相关性**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***r*** | ***df*** | ***z*** | ***p*（单尾）** |
| A | 0.025 | 114 | 0.270 | > 0.05 |
| B | -0.261 | 113 | -2.77 | < 0.01 |
| C | 0.015 | 109 | 0.154 | > 0.05 |

**注: A无眼动组，B为虚假眼动组，C为真实眼动组。**

为了解释为何真实眼动组与虚假眼动组相较于无眼动组信效度更差，我们比较了三组的整体作答时长。在比较过程中，我们剔除了总体市场超过2000s的作答时长异常问卷后，发现三组作答时长有显著差异（F(2,337) = 11.622, *p* < 0.001）。进一步进行多重比较，发现三组的平均值差值如表4-10所示，即C组的平均作答时长显著高于另外两组。C组被试可能会耗费更多的认知资源在眼动测量上，从而导致其分配在完成问卷上的精力不足。

**表4-10　各组的平均总作答时长差值及其显著性**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **平均值差值** | ***SE*** | ***p*** |
| A | B | 6.081 | 20.869 | >0.05 |
|  | C | -84.421 | 21.008 | <0.01 |
| B | A | -6.081 | 20.869 | >0.05 |
|  | C | -90.503 | 80.916 | <0.01 |
| C | A | 84.421 | 21.008 | <0.01 |
|  | B | 90.503 | 20.916 | <0.01 |

**注: A无眼动组，B为虚假眼动组，C为真实眼动组。**

# 5 综合讨论

研究结果表明，学科性别刻板印象问卷本身信度较好，效度较差。被试总体上存在学科性别刻板印象且各组间无显著差异，并总体认为理科更偏向男性，文科更偏向女性。

实验一在一定程度上说明了在样本量不充足时，根据反应时、甄别题剔除不认真的、态度防御的被试会降低问卷的信度。这可能由本实验中样本量较小、问卷题量较小导致的。由克伦巴赫系数计算公式可知，该系数与样本量与题量相关度高，样本量与题量小时，该系数也会较小。当实验样本量足够大时（一般为问卷题量的5 ~ 10倍），剔除少量样本量对信度造成的影响可以忽略不计（吴明隆，2010）。然而在本实验中，由于时间与经费限制，我们的样本量仅为题目的4倍左右。在根据反应时剔除被试时，由于被试量较小，被试阅读题目速度等差异较大，难以构成正态分布，我们很难通过判断离群值的方式确定如何剔除被试，所以选用了Huang等在2012年提出的2500毫秒作为标准。若被试量增加，则可以根据被试的反应时分布得出更适合本问卷题目的反应时筛选标准，可能可以达到更好的效果。在根据甄别题剔除被试时，相较于不剔除任何被试，样本量进一步减小了，这可能是造成有甄别题组问卷信度小于无甄别题组的原因。所以，在样本量不充足时，使用绝对的反应时标准、认真程度甄别题和测谎甄别题进行数据清洗并不能如先前文献所示地提升问卷信效度。在未来进一步研究时，应扩大样本量至题量的5倍以上，预期可能会取得不同的结果。

效度方面，本实验以IAT实验*d*值绝对值为效标。从结果上看，采用认真程度甄别题（高/低频题及指定选项题）剔除被试可以显著提高问卷效标效度，而采用态度防御题及测谎题剔除被试则并不能显著提升问卷效标效度。但实际上，在采用态度防御题及测谎题剔除被试时，剔除样本量超过30%，由于效度与样本量有关，样本量小可能导致效度降低，所以并不能证明根据态度防御题与测谎题不能提升问卷效度。在未来进一步研究时，应进一步扩大样本量，从而更规范地检验根据态度防御题与测谎题剔除被试是否可以提升效标效度。

实验二的结果表明，虚假眼动可能显著降低问卷的效度与信度，这可能是由于被试行为会受到观测者的影响（Sedgwick & Greenwood, 2015）。研究表明，适度的监视可以提高员工的工作表现，因为它可以增加工作纪律和责任感（Davidson & Henderson, 2000; Holman等, 2002）。虽然在虚假眼动组，被试的眼动数据实际上并未被采集，然而，被试在被告知其眼动数据会被记录时，错误地认为自己的答题是在更严密的监控下完成的，当个人信念与社会准则相冲突时，在认为自身处于监视状态下时，信念偏差越强的被试越会表现出虚假作答的行为。

除虚假眼动组外，真实眼动组问卷的信度也降低了。这可能是由以下三个原因导致的：第一，真实眼动组实验前，必须通过Credamo平台的眼动校准测验，通过对无效被试进行访谈得知，Credamo平台的眼动校准测验通过率较低，对被试的注视点要求高。许多被试想要参与实验，但无法通过平台的眼动校准测验，故在取样时并不随机，可能存在抽样代表性误差。第二，真实眼动组在进入正式实验时，可能分配认知资源用于眼动管理，这可能导致分配给问卷作答的认知资源不足，从而导致问卷信效度降低。过往研究表明，人的认知资源是有限的，当同时进行多个任务时，需要合理分配认知资源才能胜任所有任务（Rubinstein等, 2001）。当所有任务所需认知资源总和多于被试的认知资源时，任务就会失败（Stigchel等, 2009）。在Credamo平台上，眼动实验校准时显示在屏幕上的注视点可能因为平台眼动精度不足而发生颤抖，这会使被试分配更多精力在控制眼动点上，从而使其分配在作答问卷上的认知资源不足。第三，真实眼动组进入实验前需要进行眼动校准，导致被试的问卷填写时长显著高于无眼动组与虚假眼动组，可能导致被试在填写问卷时注意力不够集中，导致问卷信效度降低。这样的结果表明，虽然目前部分平台推出了问卷眼动监测功能，然而收集在线问卷时，采用眼动反而会降低问卷信效度，而由于目前技术限制，在线眼动测量结果并不能为我们清洗数据提供有效的数据，所以目前阶段，非必要情况下，不建议在问卷作答过程中加入眼动检测。

# 6 结论

采用认真程度甄别题剔除异常被试数据可能提升问卷效度，在样本量不充足时，采用绝对反应时标准、认真程度甄别题、态度防御与测谎甄别题清洗被试数据反而会降低问卷信度。采用虚假眼动可能降低问卷信效度。在目前的在线眼动精度不足的现实条件下，采用真实眼动可能降低问卷信度。

# 7 备注

## 7.1 备注1

由于Credamo平台功能限制，内隐联想测验（IAT）中的**错误惩罚计时无法实现**，因此，我们参考过往文献，将错误惩罚改为给错误试次增加 500ms反应时。

## 7.2 备注2

在此研究中，由于平台功能限制，我们使用了多份协作问卷，以满足研究需要。具体原因在于我们计划在在线问卷中引入眼动功能，然而，目前所采用的Credemo平台**仅支持每份问卷插入一个行为实验模块**。为了达到研究目的，我们**不得不使用多份问卷**。为了确保被试的参与体验，我们将多份协作问卷整合在一起，以确保每个被试仅需完成一份问卷。

实验一，无甄别，CI： https://www.credamo.com/u/bBKQXynkl4X

实验一，无甄别，IC： https://www.credamo.com/u/opw5RBzbmR0

实验一，认真作答甄别，CI： https://www.credamo.com/u/g4K2y6b6Rwg

实验一，认真作答甄别，IC： https://www.credamo.com/u/WnKxEwQbzNB

实验一，说谎甄别，CI： https://www.credamo.com/u/vP21vKmArWQ

实验一，说谎甄别，IC： https://www.credamo.com/u/bBKQXAV8RaJ

实验二，无眼动，CI： https://www.credamo.com/u/NEK4vo4d2WJ

实验二，无眼动，IC： https://www.credamo.com/u/laYj4Byzg49

实验二，虚假眼动，CI： https://www.credamo.com/u/XEKvGwol12G

实验二，虚假眼动，IC： https://www.credamo.com/u/XEKvGm3l3xd

实验二，真实眼动，CI： https://www.credamo.com/u/opw5RqXNkgW

实验二，真实眼动，IC：https://www.credamo.com/u/opw5Rwj5a60

# 8 参考文献

吴明隆.(2010).问卷统计分析实务：SPSS操作与应用. 重庆: 重庆大学出版社.

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, *29*, 497–519. https://doi.org/10.1093/ijpor/edw007

Barge, S., & Gehlbach, H. (2012). Using the Theory of Satisficing to Evaluate the Quality of Survey Data. *Research in Higher Education*, *53*(2), 182–200. https://doi.org/10.1007/s11162-011-9251-2

Clark, J. W., & Stephane, A. L. (2018). Affordable Eye Tracking for Informed Web Design. 收入 A. Marcus & W. Wang (编), *DESIGN, USER EXPERIENCE, AND USABILITY: THEORY AND PRACTICE, DUXU 2018, PT I* (卷 10918, 页 346–355). Springer International Publishing Ag. https://doi.org/10.1007/978-3-319-91797-9\_24

Colledani, D., & Ciani, A. C. (2021). A Worldwide Internet Study Based on Implicit Association Test Revealed a Higher Prevalence of Adult Males’ Androphilia than Ever Reported Before. *JOURNAL OF SEXUAL MEDICINE*, *18*(1), 4–16. https://doi.org/10.1016/j.jsxm.2020.09.011

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Davidson, R., & Henderson, R. (2000). Electronic Performance Monitoring: A Laboratory Investigation of the Influence of Monitoring and Difficulty on Task Performance, Mood State, and Self-Reported Stress Levels. *Journal of Applied Social Psychology*, *30*(5), 906–920. https://doi.org/10.1111/j.1559-1816.2000.tb02502.x

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The Differential Impacts of Two Forms of Insufficient Effort Responding. *Applied Psychology*, *67*(2), 309–338. https://doi.org/10.1111/apps.12117

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. https://doi.org/10.1007/s10869-016-9479-0

Francavilla, N. M., Meade, A. W., & Young, A. L. (2019). Social Interaction and Internet-Based Surveys: Examining the Effects of Virtual and In-Person Proctors on Careless Response: Proctors and Careless Response. *Applied Psychology*, *68*(2), 223–249. https://doi.org/10.1111/apps.12159

Grau, I., Ebbeler, C., & Banse, R. (2019). Cultural Differences in Careless Responding. *Journal of Cross-Cultural Psychology*, *50*(3), 336–357. https://doi.org/10.1177/0022022119827379

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Havan, P., & Kohut, M. (2023). Details Matter: The Effect of Different Instructions and Their Order on the Bias of Measured Personality Traits by Social Desirability. *STUDIA PSYCHOLOGICA*, *65*(2), 154–164. https://doi.org/10.31577/sp.2023.02.872

Holman, D., Chissick, C., & Totterdell, P. (2002). The Effects of Performance Monitoring on Emotional Labor and Well-Being in Call Centers. *Motivation and Emotion*, *26*(1), 57–81. https://doi.org/10.1023/A:1015194108376

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. https://doi.org/10.1007/s10869-011-9231-8

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Kam, C. C. S. (2019). Careless Responding Threatens Factorial Analytic Results and Construct Validity of Personality Measure. *FRONTIERS IN PSYCHOLOGY*, *10*, 1258. https://doi.org/10.3389/fpsyg.2019.01258

Marini, M., Waterman, P. D., Breedlove, E., Chen, J. T., Testa, C., Reisner, S. L., Pardee, D. J., Mayer, K. H., & Krieger, N. (2021). The target/perpetrator brief-implicit association test (B-IAT): An implicit instrument for efficiently measuring discrimination based on race/ethnicity, sex, gender identity, sexual orientation, weight, and age. *BMC Public Health*, *21*(1), 158. https://doi.org/10.1186/s12889-021-10171-7

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*, *84*, 79–83. https://doi.org/10.1016/j.paid.2014.08.021

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

Nayak, M., & K A, N. (2019). *Strengths and Weakness of Online Surveys*. *24*, 31–38. https://doi.org/10.9790/0837-2405053138

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology. Human Perception and Performance*, *27*(4), 763–797. https://doi.org/10.1037//0096-1523.27.4.763

Sedgwick, P., & Greenwood, N. (2015). Understanding the Hawthorne effect. *BMJ (online)*, *351*, h4672. https://doi.org/10.1136/bmj.h4672

Sohn, M. N., McMorris, C. A., Bray, S., & McGirr, A. (2021). The death-implicit association test and suicide attempts: A systematic review and meta-analysis of discriminative and prospective utility. *PSYCHOLOGICAL MEDICINE*, *51*(11), 1789–1798. https://doi.org/10.1017/S0033291721002117

Soland, J., Wise, S. L., & Gao, L. (2019). Identifying Disengaged Survey Responses: New Evidence Using Response Time Metadata. *Applied Measurement in Education*, *32*(2), 151–165. https://doi.org/10.1080/08957347.2019.1577244

Stigchel, S. V. der, Belopolsky, A. V., Peters, J. C., Wijnen, J. G., Meeter, M., & Theeuwes, J. (2009). The limits of top-down control of visual attention. *Acta Psychologica*, *132*(3), 201–212. https://doi.org/10.1016/j.actpsy.2009.07.001

Toepoel, V., & Emerson, H. (2017). Using experts’ consensus (the Delphi method) to evaluate weighting techniques in web surveys not based on probability schemes. *Mathematical Population Studies*, *24*(3), 161–171. https://doi.org/10.1080/08898480.2017.1330012

Wright, K. (2000). Perceptions of on‐line support providers: An examination of perceived homophily, source credibility, communication and social support within on‐line support groups. *Communication Quarterly*, *48*(1), 44–59. https://doi.org/10.1080/01463370009385579

Wright, K. B. (2005). Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, *10*(3), JCMC1034. https://doi.org/10.1111/j.1083-6101.2005.tb00259.x

Zhong X., Li M., & Li L. (2021). Preventing and detecting insufficient effort survey responding. *Advances in Psychological Science*, *29*(2), 225–237. https://doi.org/10.3724/SP.J.1042.2021.00225

Zou, J., & Zhang, H. (2019). New Key Point Detection Technology Under Real-Time Eye Tracking. *MECHATRONIC SYSTEMS AND CONTROL*, *47*(2), 71–76. https://doi.org/10.2316/J.2019.201-2969

**ABSTRACT**

**PURPOSE** The primary aim of this research was to intensively investigate strategies to bolster the reliability of online questionnaires. The inclination of online questionnaire participants to respond either carelessly or dishonestly, for various reasons, poses potential threats to the questionnaire's validity and reliability.

**METHODS** We employed the Credamo online experimental platform for questionnaire distribution and subsequent experiments. We scrutinized three distinct data cleaning techniques to ascertain their effectiveness in enhancing questionnaire reliability: discriminative questions, response time analysis, and eye-tracking measures. The Implicit Association Test was utilized as the criterion measure to evaluate the potential influence of these data cleaning methods on the questionnaire's criterion validity.

**RESULTS** The study unveiled that the removal of data with atypical response times significantly improved the questionnaire's reliability. Implementing specifically crafted discriminative questions to discard outlier participant responses considerably elevated the questionnaire's validity. Interestingly, real-time online eye-tracking seemed to diminish the questionnaire's validity and reliability, whereas the use of simulated eye-tracking notably augmented its reliability.

**附录**

**附1 分类词**

**女性：**女孩、阿姨、女儿、妻子、妇女、母亲、奶奶

**男性：**男人、男孩、爸爸、爷爷、丈夫、儿子、叔叔

**文科：**哲学、人类学、艺术、文学、英语、音乐、历史

**理科：**生物学、物理学、化学、数学、地质学、天文学、工程学

**附2 学科性别刻板印象问卷**

|  |  |  |
| --- | --- | --- |
| **题目（李克特五点量表）** | | **维度** |
| 1 | 你是否认为某些性别在某些学科领域的贡献被低估或忽视了？ | 创新和研究 |
| 2 | 你是否认为性别在学科研究方向的选择上会有所限制？ | 创新和研究 |
| 3 | 你是否认为某个性别在某些学科领域内更具创新能力？ | 创新和研究 |
| 4 | 你是否认为某些学科领域内的研究方向适合某一性别？ | 创新和研究 |
| 5 | 你是否感觉在某些学科领域内，某个性别的贡献更受重视？ | 创新和研究 |
| 6 | 你是否认为在学科中的领导地位上女性的代表性不足？ | 领导和代表性 |
| 7 | 你是否认为女性在学科领域的领导地位上面临更多挑战？ | 领导和代表性 |
| 8 | 你是否认为学科领域内的领导者应该更多地代表不同性别？ | 领导和代表性 |
| 9 | 你是否认为某些学科领域内的领导地位应该主要由某个性别担任？ | 领导和代表性 |
| 10 | 你认为性别刻板印象是否影响了学生在某些学科中的自信？ | 性别的成绩差异 |
| 11 | 你是否感觉不同性别的学生在某些学科的学业表现上存在差异？ | 性别的成绩差异 |
| 12 | 你是否觉得在某些学科中，某个性别的能力更强？ | 性别的成绩差异 |
| 13 | 你是否觉得在某些学科中，某个性别更具有优势？ | 性别的成绩差异 |
| 14 | 你是否认为某个性别具有学习某些学科的天赋？ | 性别的成绩差异 |
| 15 | 你是否觉得某些学科是某个性别的领域？ | 学科选择和倾向 |
| 16 | 你是否认为某些学科更适合某个性别？ | 学科选择和倾向 |
| 17 | 你是否认为在某些学科中，性别会限制人的选择和发展？ | 学科选择和倾向 |
| 18 | 你是否认为社会对性别的期望会影响学科的选择？ | 学科选择和倾向 |
| 19 | 你认为性别刻板印象是否影响了学生在某些学科中的兴趣？ | 学科选择和倾向 |
| 20 | 你是否感觉在某些学科中，某一性别学生更受鼓励和支持？ | 学科选择和倾向 |
| 甄1 | 你认识的所有女生在文科考试中都能取得满分？ | 认真甄别题-低频 |
| 甄2 | 你认为有的时候男生在理科考试中的成绩会比女生差？ | 认真甄别题-高频 |
| 甄3 | 本题请选择2。 | 认真甄别题-固定 |
| 甄4 | 我所认识的同性里不是个个我都喜欢。 | 说谎甄别题 |
| 甄5 | 偶尔我会想到一些坏得说不出口的话。 | 说谎甄别题 |
| 甄6 | 有时我也会对另一性别产生某些偏见。 | 说谎甄别题 |