

首届全国大学生心理与行为在线实验精英赛

（校内赛· 研究设计）

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 <input checked="" type="checkbox"/> B 揭榜赛道 <input type="checkbox"/>
问卷分享链接 (Credamo 见 数平台)	作答链接（被试所见）： <a href="https://www.credamo.com/s/zMBZvy">https://www.credamo.com/s/zMBZvy</a> <a href="https://www.credamo.com/s/q6nqyq">https://www.credamo.com/s/q6nqyq</a> 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度的方法。在线问卷具有多种优势，但也面临着数据污染的风险，因为参与者可能不认真或不诚实地回答问题，从而降低问卷的信效度。为了解决这一问题，本研究使用了 Credamo 见数在线实验平台来发布问卷和实验，以探索采用甄别题、答题反应时以及眼动数据三种方式对问卷数据进行清洗是否能够增强问卷的信度。此外，还对在问卷中设置虚假眼动是否也能提高问卷的信度进行了进一步探讨。本研究将以内隐联合测验结果为效标，来评估不同数据清洗方法对问卷信度的影响。

**【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、答题反应时与眼动数据清洗问卷数据是否能够提升问卷的信度。本研究中的问卷与实验将在 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上出口的话”。这类题目旨在检测被试在回答问卷时是否如实回答。

## 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種族的正面态度，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如何。

## 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度？
- 利用眼动清洗数据是否能够提升在线问卷信度？问卷中设置虚假眼动是否也能够提升在线问卷信度？若能，其作用与利用眼动清洗数据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度。

### 4.1.2 被试

利用 Credamo 平台收集 120 名中国本土长大的、汉语为母语的 18~25 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  $2 \times 3$  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、采用说谎甄别题。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 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。



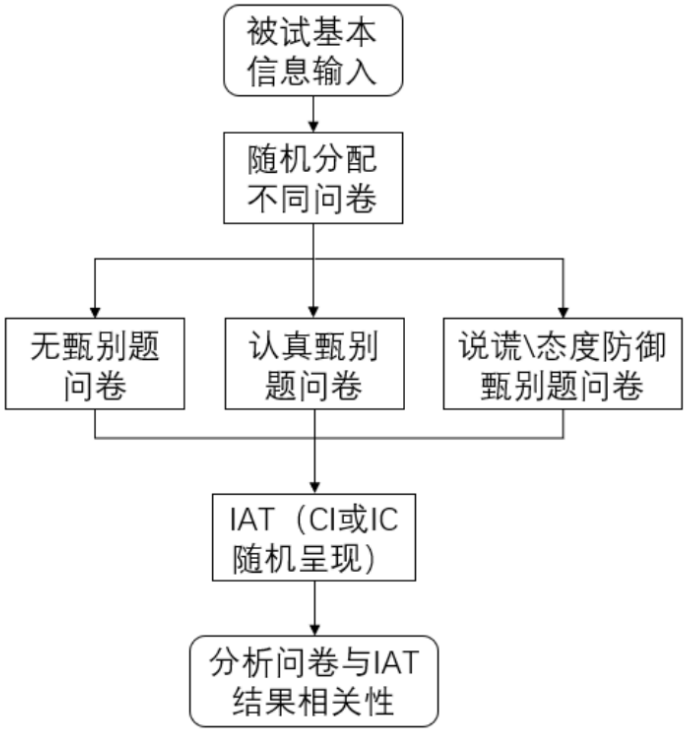


图 4-1 利用甄别题与反应时清洗数据对在线问卷信度的影响研究流程

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。

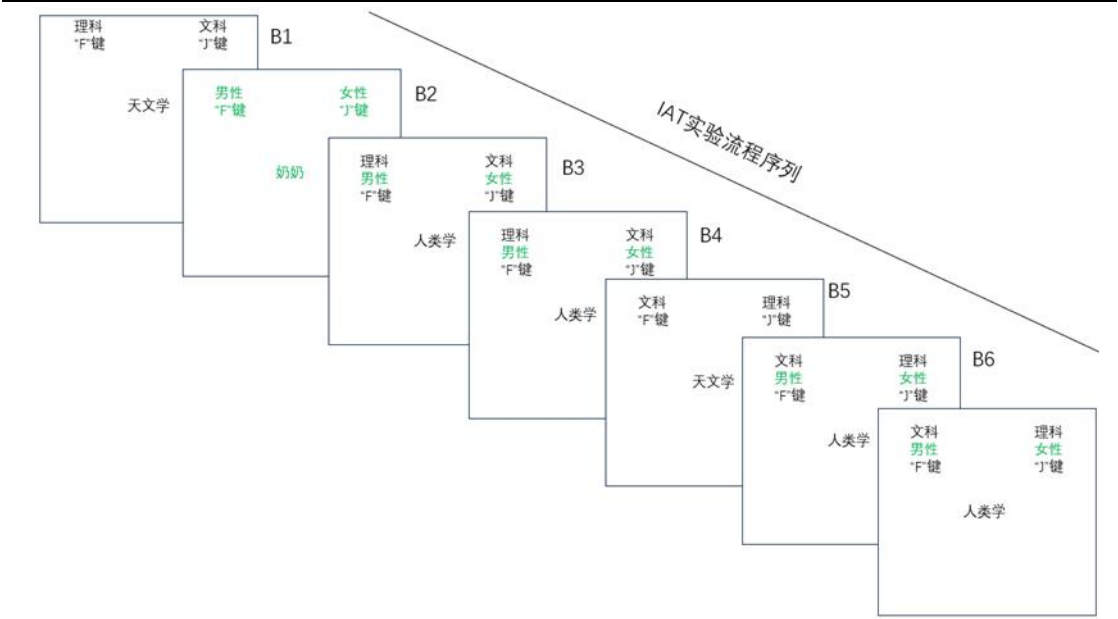


图 4-2 内隐联想测验（IAT）流程序列

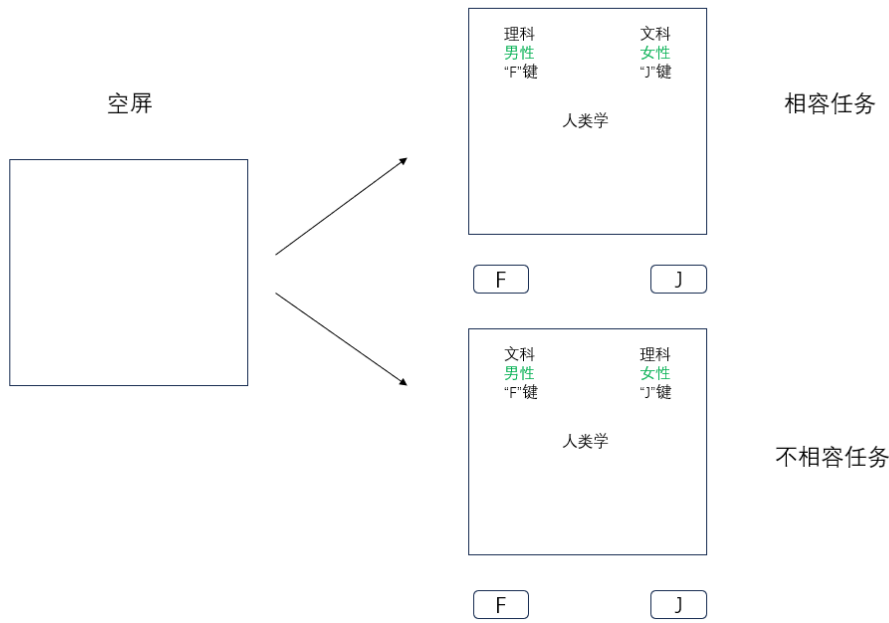


图 4-3 内隐联想测验（IAT）单次实验流程

4.1.6 预期实验结果

预期收集反应时组问卷信度高于不收集组，预期采用甄别题组问卷信度高于无甄别题组，且采用不同类型甄别题的两组问卷信度有显著差异。

## 4.2 研究二：采用眼动对在线问卷信度的影响

### 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度。
2. 探究采用虚假眼动是否可以提升在线问卷的信度。

### 4.2.2 被试

利用 Credamo 平台收集 90 名中国本土长大的、汉语为母语的 19~23 岁本科生被试，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 4-4 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 具体过程同研究一。

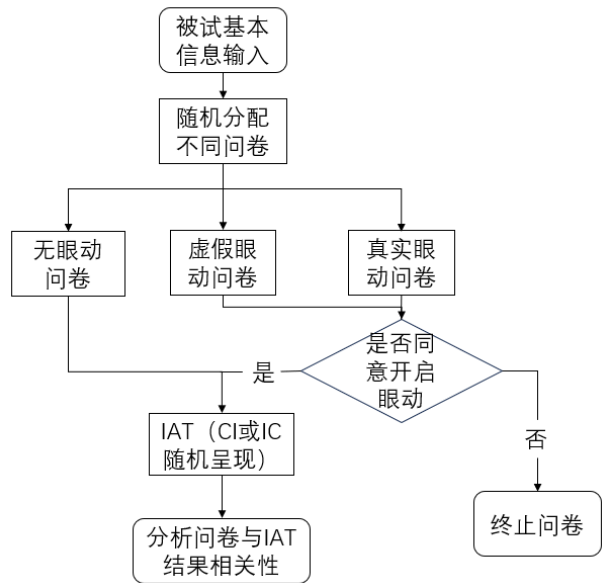


图 4-4 采用眼动对在线问卷信度的影响研究流程

#### 4.2.6 预期结果

预期真实眼动组问卷信度高于无眼动组与虚假眼动组，虚假眼动组问卷信度高于无眼动组。具体表现为真实眼动组的问卷得分与 IAT 实验中得出的效应量之间的相关显著高于另两组，且虚假眼动组的问卷得分与 IAT 效应量的相关显著高于无眼动组。

## 5 备注

### 5.1 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。

### 5.2 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/WnqJGw3nZg4>

实验一，无甄别，IC: <https://www.credamo.com/u/XEgoj8djJWw>

实验一，认真作答甄别，CI: <https://www.credamo.com/u/Q3rqvwr1a8G>

实验一，认真作答甄别，IC: <https://www.credamo.com/u/bBOJqypr3Dn>

实验一，说谎甄别，CI: <https://www.credamo.com/u/la15ORoWW82>

实验一，说谎甄别，IC: <https://www.credamo.com/u/la15OBOKNAX>

实验二，无眼动，CI: <https://www.credamo.com/u/oprd1wD9kEd>

实验二，无眼动，IC: <https://www.credamo.com/u/mj3K9dYgm6g>

实验二，虚假眼动，CI: <https://www.credamo.com/u/WnqJoJoXm3B>

实验二，虚假眼动，IC: <https://www.credamo.com/u/la15zAOdKdg>

实验二，真实眼动，CI: <https://www.credamo.com/u/la15z0BvvB2>

实验二，真实眼动，IC: <https://www.credamo.com/u/WnqJoxzY4YE>

## 6 参考文献

- 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](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>
- 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>
- 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>
- 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>
- 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>



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

首届全国大学生心理与行为在线实验精英赛（校内赛 • 研究设计）

---

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