

Running Human Subject Experiments via Online Crowdsourcing

BIBC2025 talk

Patrick Li

RSFAS, ANU



About me

Hi! I am Patrick Li.

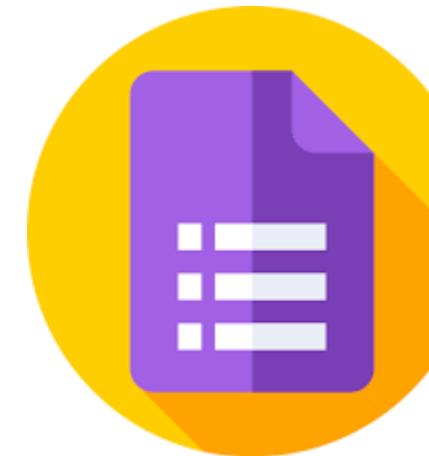
- I hold a PhD in Statistics. My research focused on computer vision and data visualization, with an emphasis on developing visual analytics methods to assess residual plots.
- I am currently working at ANU for the AAGI project, primarily on machine learning, image analytics, and plant phenotyping projects.

Online human subject experiments

Online Human-Subject Experiments are research studies in which human participants engage in tasks, surveys, or behavioral assessments **over the internet**.

Modern online platforms make it easier than ever to run human-subject experiments.

- **Ease of use:** No need to physically gather participants in a lab.
- **Large-scale data collection:** broad and diverse participant pool across different regions.
- **Rapid deployment and iteration:** Experiments can be launched and modified at any time.



Designing online experiments

Online experiments are similar to lab experiments but require additional considerations:

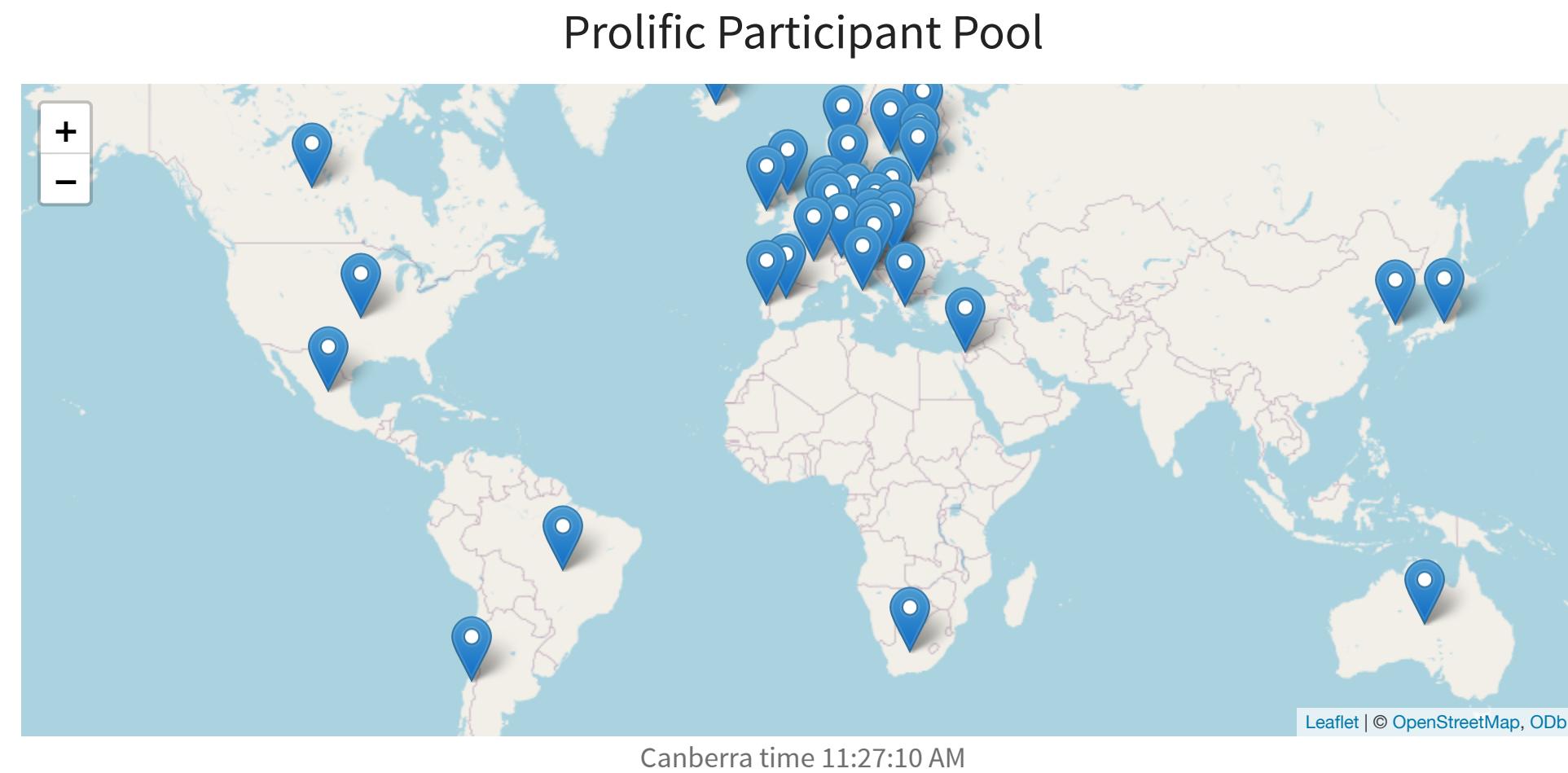
- Participant diversity
- Inconsistent responses
- Lower engagement
- Limited participant information
- Pilot testing
- Data quality control

Participant diversity

Online participants are highly heterogeneous.

Consider variations in:

- Gender, age, culture, education, geographic region
- Prior exposure to similar tasks



Inconsistent responses

Participants may respond inconsistently, meaning they provide **different answers to the same or highly similar questions under identical conditions.**

Possible reasons include:

- Multitasking or environmental distractions
- Cognitive fatigue or overload
- Ambiguity in the question wording
- Changes in interpretation or internal state between responses

Inconsistent responses can reduce the reliability of your data and should be considered when designing surveys or experiments.

Engagement & Task Design

Online participants often approach studies as **short-term tasks**, which can lead to **low engagement**. Designing tasks with participant experience in mind helps maintain attention and improve data quality.

- Keep tasks **short and focused**
- Use a **clear interface** and ensure **smooth task flow**
- Consider **higher compensation** for longer or more demanding tasks

The image shows a mobile application interface. At the top, a question is displayed in a white box with a thin gray border: "What is the color of the sky on a clear day?". Below the question are four options, each in its own rectangular box with rounded corners and a thin gray border:

- A) Blue
- B) Green
- C) Red
- D) Yellow

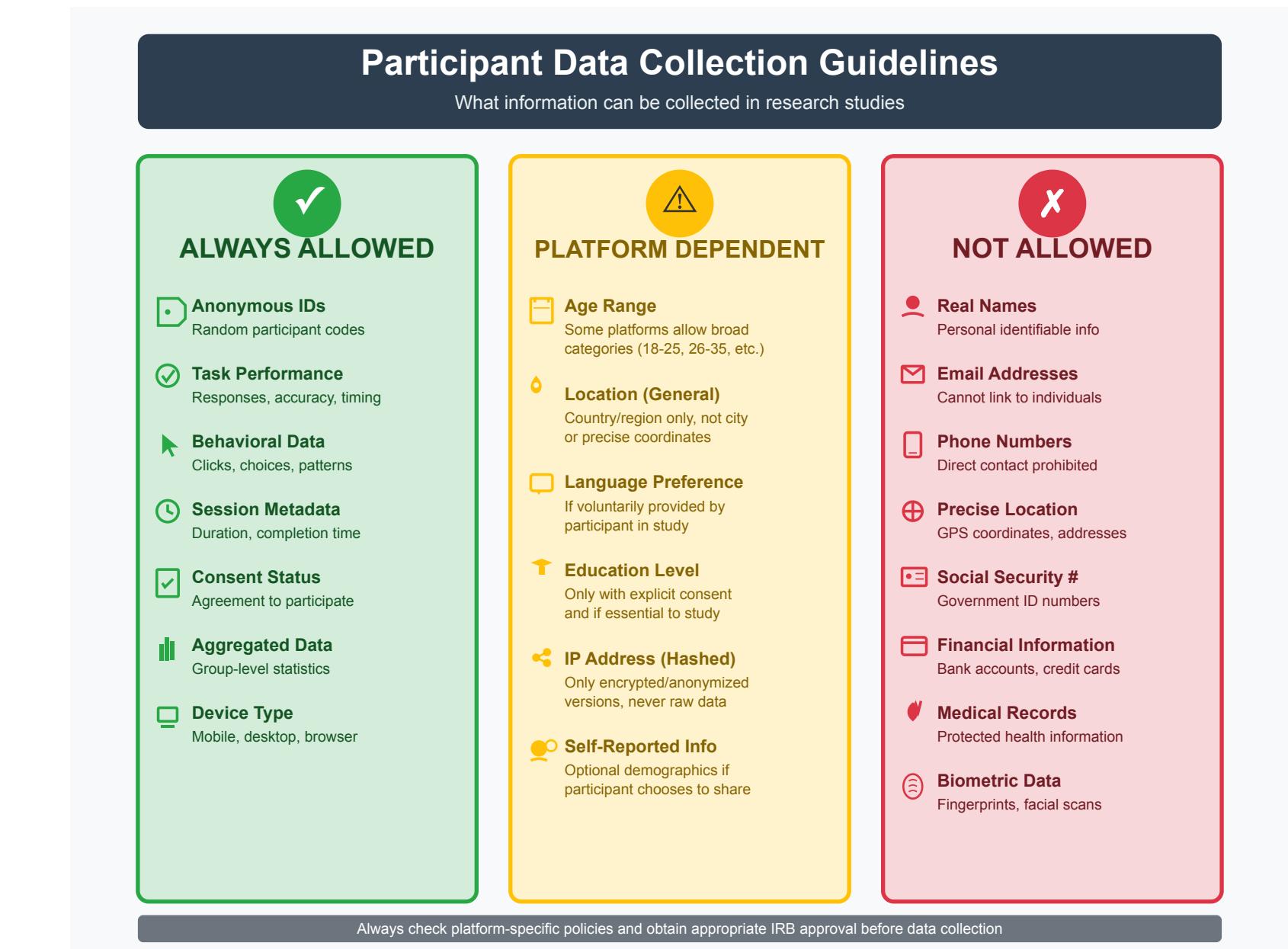
At the bottom of the screen, there is a horizontal progress bar consisting of a green segment followed by a gray segment. At the very bottom center, the text "Canberra time 11:27:10 AM" is visible.



Limited participant information

Due to anonymity and ethical constraints, you typically cannot collect detailed personal data.

- Plan to collect minimal demographic data when designing your experiment.
- Ensure your analysis does not rely on attributes that cannot be legally or practically obtained.



Pilot testing

Pilot testing involves running **small-scale studies** before the main experiment to identify potential issues early.

It helps detect:

- **Design flaws**: problems in the study setup or procedure that could affect results.
- **Timing issues**: tasks that take too long or too short, affecting participant engagement.
- **Unclear instructions**: questions or tasks that participants may misinterpret.
- **Unnecessary costs**: resources spent on elements that do not contribute meaningful data.

Data quality control

Data quality control ensures that responses are **reliable, human-generated**, and suitable for analysis.

Include mechanisms to detect careless or automated responses:

- Attention checks:
 - Please select “Blue” from the options below. (A) Red (B) Blue (C) Yellow
- Repeated items:
 - identical or near-identical questions to test consistency.
- Minimum-time thresholds

Key parameters of an online experiment

An online experiment is often characterised by several core design parameters:

- Participant count (P)
- Assignments per participant (A)
- Total assignments ($P \times A$)
- Task duration (T)
- Treatment combinations (K)
- Replications (R)
- Budget considerations (B)
 - Compensation (C)
 - Platform fees (F)
 - Expected exclusions (E)

Ethics & HREC approval

All online experiments must comply with NHMRC ethical guidelines and institutional policies, and obtain HREC approval.

Low-risk research

Procedures involve no more than everyday inconvenience or mild discomfort.

Typical online surveys and behavioural tasks fall into this category.

Higher-risk research

Possibility of physical or psychological harm, distress, or exposure to sensitive content.

Includes studies with vulnerable populations or topics requiring extra safeguards.

Low-risk study requirements

Low-risk studies have a high approval rate. When applying, you typically need to provide:

1. Study description and procedures
2. Participant information and consent form
3. Risk assessment (minimal for low-risk projects)
4. Data handling and privacy plan

Ongoing studies may also require **annual review**.

Show Inactive Sections

Human Ethics Application Form

Section	Questions
Start	Start Here Amendment Details Checklist
Section A	Investigators Background and Aims Research Scope Benefits and Risks Risk Management Project Details
Section B	Participant Groups Participant Details Recruitment Reimbursement and Participant Information Informed Consent Opt-Out Consent
Section C	Data Collection Methods Data Collection Details Research Involving Radiation Participant Groups and Data Collection Methods Research Procedures Diagnostic Measures
Section H	Privacy and confidentiality Waiver of Consent Justification Collection, Use and Disclosure of Information
Section I	Data access and security
Section J	Research outcomes
Section K	Documents and Declarations

Canberra time 11:27:10 AM



Online Crowdsourcing Platforms

When choosing a platform for online studies, consider several key factors:

- Prescreening options
- Pricing & compensation
- Participant quality
- Platform policies & ethics
- Study integration

Prescreening Options

Most platforms provide tools to prescreen participants before your study begins. Proper prescreening helps ensure your sample matches your research requirements.

- **Demographics:** Age, gender, country, language, or other characteristics. Some platforms offer advanced options, such as employment status, but these may incur extra fees.
- **Performance-based filters:** Some platforms allow low-acceptance prescreening.

Be aware that time zones may introduce unintended biases.

Age	Sex	Ethnicity simplified	Country of birth	Country of residence	Nationality	Language	Student status	Employment status
20	Male	White	Portugal	Portugal	Portugal	Portuguese	Yes	Other
22	Male	White	Portugal	Portugal	Portugal	Portuguese	Yes	Unemployed (and job seeking)
23	Male	White	Poland	Poland	Poland	Polish	Yes	Due to start a new job within the next month
20	Male	White	Portugal	Portugal	Portugal	Portuguese	Yes	Due to start a new job within the next month
22	Male	Other	Argentina	Spain	Spain	Spanish	Yes	Unemployed (and job seeking)
21	Male	White	Portugal	Portugal	Portugal	Portuguese	Yes	Part-Time
22	Male	White	Portugal	Portugal	Portugal	Portuguese	Yes	Part-Time
22	Male	White	Portugal	Portugal	Portugal	Portuguese	Yes	Other
23	Male	White	Portugal	Portugal	Portugal	Portuguese	Yes	Full-Time
22	Male	White	Poland	Poland	Poland	Polish	Yes	DATA_EXPIRED
23	Male	Mixed	Portugal	Portugal	Portugal	Portuguese	Yes	Other
29	Male	White	Portugal	Portugal	Portugal	Portuguese	DATA_EXPIRED	DATA_EXPIRED
20	Male	White	Portugal	Portugal	Portugal	Portuguese	Yes	Other

Price & Compensation

Platforms vary in how they charge and compensate participants:

- Dynamic vs static rate
- Platform fee
- Minimum rate constraints
- Participant bonuses

Participant	Reward	Platform fees	Vat	Total
613add22d171e75cc0b3576e	£1.88	£0.62	£0.00	£2.50
6146385561e8f95ff4f3b5d6	£1.88	£0.62	£0.00	£2.50
615df2e8304b4a312cb5cd6f	£1.88	£0.63	£0.00	£2.51

Participant quality

It's important to consider the overall quality of participants a platform provides:

- Past reliability: Check reports, blogs, or shared experiences from other researchers to see how dependable participants are on this platform.
- Pilot studies: Use small, inexpensive pilot studies to assess whether participants engage with tasks as expected.
- Task design: Ensure your study can be completed easily and remains engaging, since overly difficult or boring tasks can reduce data quality.

Study integration

Many platforms allow you to run studies directly within their system. These typically support conventional question types such as multiple-choice, free-text, or rating scales.

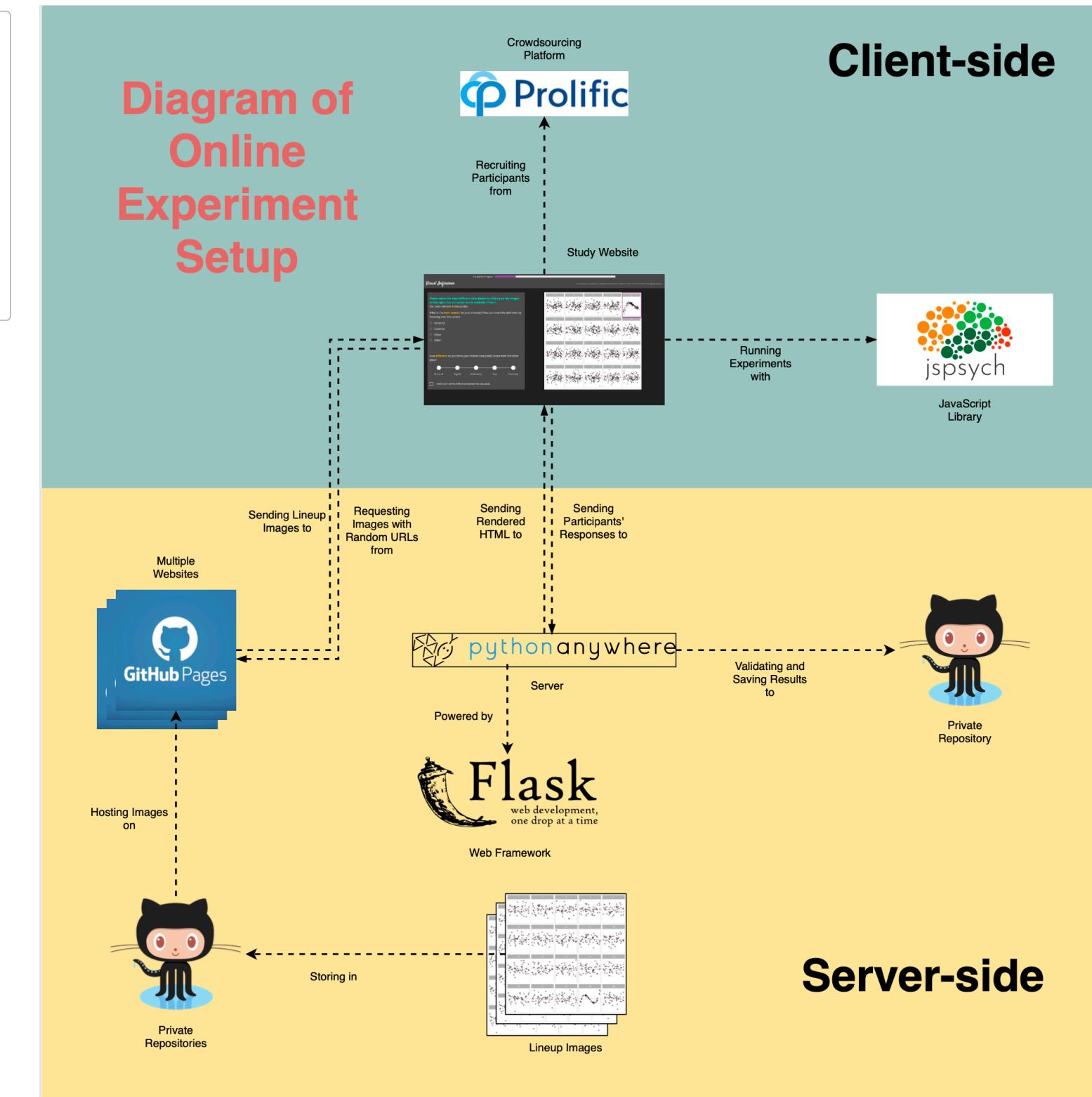
Some platforms **only allow** you link to external surveys or experiments. This often requires participants to enter a completion code. Common options include:

- External survey websites: Google Forms, SurveyMonkey, or Qualtrics.
- Custom websites or apps: For example, a Shiny app for interactive experiments.

Study: residual plot assessment

Li, W., Cook, D., Tanaka, E., & VanderPlas, S. (2024). A plot is worth a thousand tests: Assessing residual diagnostics with the lineup protocol. *Journal of Computational and Graphical Statistics*, 33(4), 1497-1511.

- Recruited participants via **Prolific** and run experiments externally.
- Experiment requires interactive features (e.g., clicking on images), built custom website using **Python Flask + jsPsych**, hosted on PythonAnywhere. Responses validated by server and synced to private GitHub repo.





Australian
National
University

Canberra time 11:27:10 AM

Study: residual plot assessment

Participant behavior varies:

- 5% of participants hang on the study for a long time before submitting, which could be considered **potentially abusing platform rules** and can skew median completion times and platform costs.
 - **10%** respond randomly and fail attention checks.
-
- Prolific applies a **33% platform fee!**
 - Complex designs can lead to unexpected issues (we lost ~AUD 1,000).
 - **Run small proof-of-concept experiments** before full deployment.

Overall, documentation is solid and the UI is acceptable, and participant quality is moderate, though dynamic pricing can be tricky for fixed budgets.

Study: food safety survey

Debias officer judgments on food safety images/videos from India by establishing ground truth using crowdsourcing. Not yet published, conducted with Klaus Ackermann and Denni Tommasi.

- Participants answer **MCQs** based on images or videos.
- Used **MTurk** with existing survey templates (pure HTML), customized for this study.
- CSV upload contains metadata (e.g., image/video IDs) linked to Amazon S3.



Question 1

Is there a food preparation area? That is, a dedicated space for preparing food?

- Yes No I don't know

Question 2

Is the top area (e.g. counter top) for food preparation waterproof?

- Yes No I don't know

Question 3

Is the top area for food preparation (e.g. a counter top) cracked or does have holes?

- Yes No I don't know

Study: food safety survey

- MTurk recruitment is extremely fast, ~500 participants in 5–10 minutes.
- Can prescreen participants in India; more advanced prescreening costs extra.
- Number of replications can be easily set.
- Data is stored on the platform and downloadable as CSV, including rich participant metadata.
- Task is short (1–2 minutes), relatively simple, and quality is reasonable with high consistency in answers.

Thanks! Any questions?

-  TengMCing
-  patrick.li@anu.edu.au
-  <https://patrickli.org>