DataPruebas: An Online Platform for Data Collection

Gustavo E. Juantorena (0000-0001-8248-7630, gjuantorena@gmail.com)[1,+],
Lara Gauder (0000-0001-6242-1546, mgauder@dc.uba.ar)[1,+],
Pablo Laciana (0009-0006-9286-8852, placiana@gmail.com)[2],
Luciana Ferrer (0000-0002-0426-8683, lferrer@dc.uba.ar)[1],
Juan E. Kamienkowski (0000-0002-5725-6539, juank@dc.uba.ar)[1,3,4]

[1] *Laboratorio de Inteligencia Artificial Aplicada, Instituto de Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires - CONICET, Argentina;*

[2] *Instituto de Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires - CONICET, Argentina*

[3] *Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina*

[4] *Maestría en explotación de datos y descubrimiento del conocimiento, FCEyN-FI, UBA*

[+] *Both authors equally contributed to the present work*

Abstract. We introduce DataPruebas, a user-friendly platform developed in Spanish, designed to streamline scientific data collection. The platform leverages public interest in contributing to scientific research. Participants can register, browse, and participate in various experiments either online with instant access or asynchronously through a flexible scheduling system.

For researchers, the platform simplifies the process of uploading experimental protocols and specifying participant criteria based on demographic variables such as age, place of birth, and other relevant characteristics. It is compatible with multiple open-source experiment frameworks, including PsychoPy, jsPsych, and lab.js, allowing researchers to integrate existing paradigms with minimal adaptation. This flexibility makes it easy to migrate previously developed experiments or design new ones using familiar tools. The platform has already been utilised to collect data on decision-making, attention, memory, and survey-based studies, among others.

In addition to supporting traditional behavioral experiments, DataPruebas is designed to facilitate the creation of high-quality datasets for machine learning research. It enables the collection of diverse data modalities, including audio recordings, webcam-based eye tracking, and mouse trajectories, allowing researchers to build multimodal datasets tailored to real-world applications.

Keywords: online cognitive experiments, dataset collection, community research.

1 Introduction

Reliable data acquisition plays a fundamental role in research across various disciplines, including cognitive science and artificial intelligence. This potential was demonstrated by the pioneering work of Luis von Ahn, creator of *reCAPTCHA* (von Ahn et al., 2008), *Duolingo* (von Ahn, 2013), and the concept of *games with a purpose* (Ahn, 2006). His research showed how mass online participation can be harnessed to solve computational problems and collect meaningful data at scale.

Crowdsourcing has emerged as a powerful model for collecting scientific data at scale, enabling researchers to reach diverse participant pools through the Internet. Several platforms exist, each with its unique strengths and limitations. In terms of cognitive experiments, Pavlovia (Bridges et al., 2020) supports PsychoPy-based studies but charges per participant and manages data externally. Gorilla (Anwyl-Irvine et al., 2020) offers an intuitive experiment builder with integrated hosting, though it requires a paid subscription. Cognition (Vidal, 2020) supports jsPsych-based studies with flexibility, though it lacks built-in participant management. Free options, such as JATOS (Lange et al., 2015) and IBEX (Schwarz & Zehr, 2021) also exist: JATOS offers full control but requires self-hosting and infrastructure management, while IBEX avoids self-hosting but is limited by a custom mini-language that restricts flexibility and modern integration.
On the other hand, platforms such as Amazon Mechanical Turk (Amazon Web Services, Inc.) and Scale AI (Scale AI, Inc.) are widely used for large-scale data annotation tasks, particularly in machine learning. They are effective at distributing microtasks, such as image labelling, audio transcription, or text classification, to a broad crowd of workers. However, these platforms come with important limitations: they offer little to no experimental control, lack long-term participant tracking, and provide only basic or limited demographic filtering.

While these tools have enabled significant progress in web-based research, many remain costly or technically demanding and are primarily designed for English-speaking participants, further limiting accessibility and inclusivity.

2 What is DataPruebas?

We developed DataPruebas (https://datapruebas.org/) as an open-access platform to streamline the process of scientific data collection. Designed specifically for the Spanish-speaking community, the platform utilises a modern web stack (Django[1] for the backend and React[2] for the frontend). The platform provides an accessible and intuitive interface for participants to register, explore available studies, and contribute to ongoing research.

DataPruebas supports two different modes of participation: asynchronous and synchronous experiments. In the former, participants complete tasks at their own schedule, using various devices with Internet access. This format is particularly effective for reaching large and diverse samples, as it eliminates the need for real-time coordination. In contrast, synchronous experiments are scheduled sessions that involve real-time interaction between the participant and the researcher, often using video conferencing tools or built-in communication features, and also for in-person studies. This approach is more appropriate for tasks that require supervision or immediate feedback. By offering both modes, the platform accommodates a wide range of experimental designs while making participation more accessible.

3 Platform interfaces

3.1 Participant's view

Participants access the platform by creating an account using email and password and completing a brief profile that includes key demographic variables such as birthdate, gender, place of birth and residence, and education level. Figure 1 illustrates the application's login interface. The design includes an explanatory section outlining the purpose of participating in DataPruebas, along with a user-friendly layout to encourage engagement.

The registration information helps match participants with relevant studies while preserving anonymity and ensuring ethical data handling. After the subject participates in an experiment, the researchers can access participant information through an interface that displays only demographic data, excluding any personally identifiable information.

Once registered, participants can browse available experiments and filter them based on participation mode (asynchronous, completed at any time, or synchronous, requiring live interaction with a researcher). For synchronous experiments, the platform includes a calendar view that allows participants to schedule an appointment at their convenience.

[1] https://www.djangoproject.com/
[2] https://react.dev/

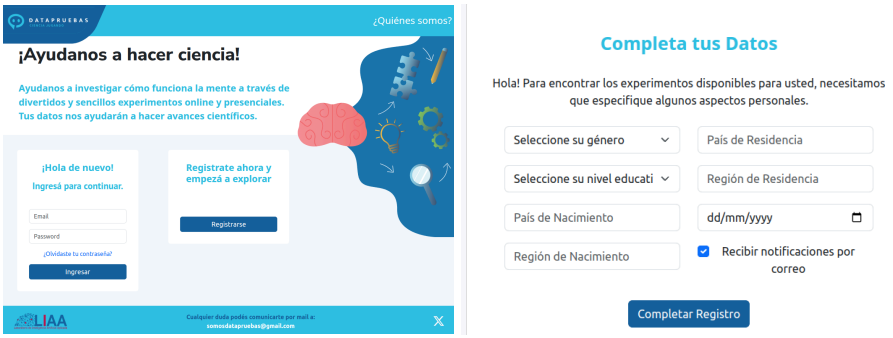

Fig. 1. On the left, we can observe DataPruebas's login screen, while on the right is the form that participants must complete when accessing the site for the first time. This enables us to access the demographic data of the participant.

To join an asynchronous study, participants select an experiment and begin the task immediately. Depending on the study's design, they may pause and return later to complete it at their convenience. In the case of synchronous studies, participation requires selecting an available time slot in advance through the platform's calendar interface. Participants must await the researcher's approval confirming their acceptance for the requested date. Upon approval, they will receive the necessary details to participate in the experiment on the scheduled date.

3.2 Researcher's view

General. For asynchronous experiments, researchers can develop their tasks using standard web technologies such as HTML, CSS, and JavaScript, or opt for specialised frameworks such as jsPsych (de Leeuw, 2015), lab.js (Henninger et al., 2022), or PsychoPy/PsychoJS (Peirce et al., 2019). The latter provides a particularly accessible entry point, allowing researchers with limited coding experience to design experiments using the PsychoPy graphical interface and have them automatically compiled into JavaScript. Since these frameworks were primarily designed to be hosted on platforms like Pavlovia, we have made a concerted effort to create easy-to-follow documentation to help researchers adapt their experiments for use with our API, whether hosted on their servers or other external platforms. Researchers can choose to upload their experiments directly to our servers or simply provide an external URL, using the participant base and built-in features of DataPruebas (such as recruitment filters, scheduling, and data tracking) without having to migrate their infrastructure fully.

Synchronous studies follow a slightly different process. Researchers need to describe their experiment and define the blocks of time available for scheduling. The platform allows them to manage appointments through a built-in calendar interface and efficiently track participant sessions.

All experiments, regardless of modality, must undergo a review and approval process from the administrators before being published. This ensures compliance with ethical standards and helps maintain the quality and reliability offered to participants.

Experiment setup. Researchers can create and configure experiments by entering metadata (title and description), uploading materials (e.g., images, consent forms), and configuring parameters such as duration, repeatability, technical requirements (e.g., webcam), and eligibility criteria (e.g., age, gender, location). Participant activity is tracked in real time through an execution panel. Raw session data (timestamps, responses, metadata) can be exported in formats like JSON or ZIP for analysis and reproducibility.

4 Current limitations and future work

While DataPruebas offers a flexible infrastructure for online experimentation and data collection, there are still some important limitations. First, the platform is centrally hosted and maintained by our team, and researchers have the option to store data outside our servers (for example, by modifying the experiment code to send data to a custom endpoint). However, this process is not yet supported through our official API, which limits the ease of integration with external databases or institutional infrastructures. Making this feature more accessible is a key goal for future development.

Second, experiment creation currently relies on programming knowledge or external tools. DataPruebas does not yet include a built-in visual interface for experiment design, which can present a barrier for researchers without technical backgrounds. However, this limitation can be partially addressed by using the PsychoPy Builder, which enables the design of experiments graphically.

Looking ahead, we plan to enhance the participant experience by introducing gamification elements, such as points based on task completion, time contributed to research and performance. These features will be used to build leaderboards and offer participants a sense of progress and recognition for their contributions to science.

Finally, while the platform is already freely available, we plan to release it as open-source software to enable broader use, adaptation, and self-hosting. This step will support transparency, encourage collaboration, and allow research communities to take ownership of the infrastructure in their local or institutional contexts.

5 Conclusions

DataPruebas represents more than just a technical solution; it is a growing community of researchers and participants committed to advancing science through accessible, scalable, and ethical data collection. With several experiments running and a steadily increasing number of registered participants, the platform already boasts a diverse range of studies, contributing to the creation of high-quality datasets.

We believe that local development is crucial. Building open-source, non-profit infrastructure in Latin America empowers researchers in underrepresented regions, reduces dependency on costly external tools, and helps foster a sustainable research ecosystem rooted in the needs and realities of our communities. This also entails the challenge of continuously maintaining and enhancing these tools.

The platform is designed with flexibility in mind: it supports popular frameworks, allows for both internal and external hosting, and provides built-in tools for participant management and data export. This balance between low technical barriers and financial accessibility makes it an ideal option for researchers seeking full control over their studies without sacrificing usability.

Looking ahead, a key challenge will be to implement mechanisms that ensure the samples collected are representative of the broader population, while also maintaining the strict anonymity of the participants. Addressing these issues is essential to maximising the scientific value and ethical integrity of the datasets on the platform.

DataPruebas is an invitation to collaborate, share, and build a more inclusive and participatory scientific future.

6 References

Amazon Web Services, Inc. (n.d.). *Amazon Mechanical Turk*. https://www.mturk.com/

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407.

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). lab.js: A free, open, online study builder. *Behavior Research Methods*, *54*(2), 556–573.

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE*, *10*(6), e0130834.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Scale AI, Inc. (n.d.). *Accelerate the Development of AI Applications | Scale AI*. https://scale.com/

Schwarz, F., & Zehr, J. (2021). Tutorial: Introduction to PCIbex – An Open-Science Platform for Online Experiments: Design, Data-Collection and Code-Sharing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/uc/item/1ng1q4c6

Vidal, J. (2020). *Cognition.* Cognition. https://www.cognition.run

von Ahn, L. (2013). Duolingo: Learn a language for free while helping to translate the web. *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, 1–2.

von Ahn, L. (2006). Games with a purpose. Computer, 39(6), 92–94. Computer.

von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, *321*(5895), 1465–1468.