# Supplementary Document

## Anchoring and Alignment: Data Factors in Part-to-Whole Visualization


### Connor Bailey and Michael Gleicher


This supplement contains further details and information about the studies conducted for *Anchoring and Alignment: Data Factors in Part-to-Whole Visualization*. This includes:

- Condition Generation: the process in which a stratified sample of stimuli were drawn from the large space of conditions
- Procedure Exhibits: a record of the online user study procedure the participants followed during their participation in the study,
- Power Analysis: the results of a power analysis we conducted based on pilot data collected before running the full study in order to determine sufficient sample sizes of participants
- Exclusions: the preregistered criteria used for excluding results and the outcome of replacing participants that failed to meet the criteria
- Models: the full model results from the GLMMs fit to the data, and the model comparison results when evaluating covariates and distributions
- Demographics: the demographic data of the participants

The preregistration of the study can be found at: `https://osf.io/e36au`. The supplementary repository containing the code and data can be found at: `https://github.com/uwgraphics/PartToWhole`.

## 1 CONDITION GENERATION

The conditions were generated through a stratified random sampling process in which stimuli were selected from groups. Each participant saw 96 stimuli with the chart type conditions being evenly split between 48 pie charts and 48 stacked bar charts. The alignment conditions were stratified such that each participant saw 24 aligned, 36 near-aligned, and 36 far-from-aligned trials to ensure sufficient sample sizes of each condition. Similarly, values estimated by the participant were stratified such that each participant saw 12 anchor values, 36 near-anchor values, and 48 far-from-anchor values. The near-anchor and near alignment conditions were values within 4 of the nearest anchor value or alignment position respectively. The distribution of the full set of stimuli across anchor distance, alignment distance, anchor category, alignment category and round by 5s can be seen in Fig. 1. With the anchor and alignment distance being the distances of the value the participant will estimate from the nearest anchor or aligned mark respectively. Round by 5s refers to the distance of the part being estimated to the nearest number divisible by 5.

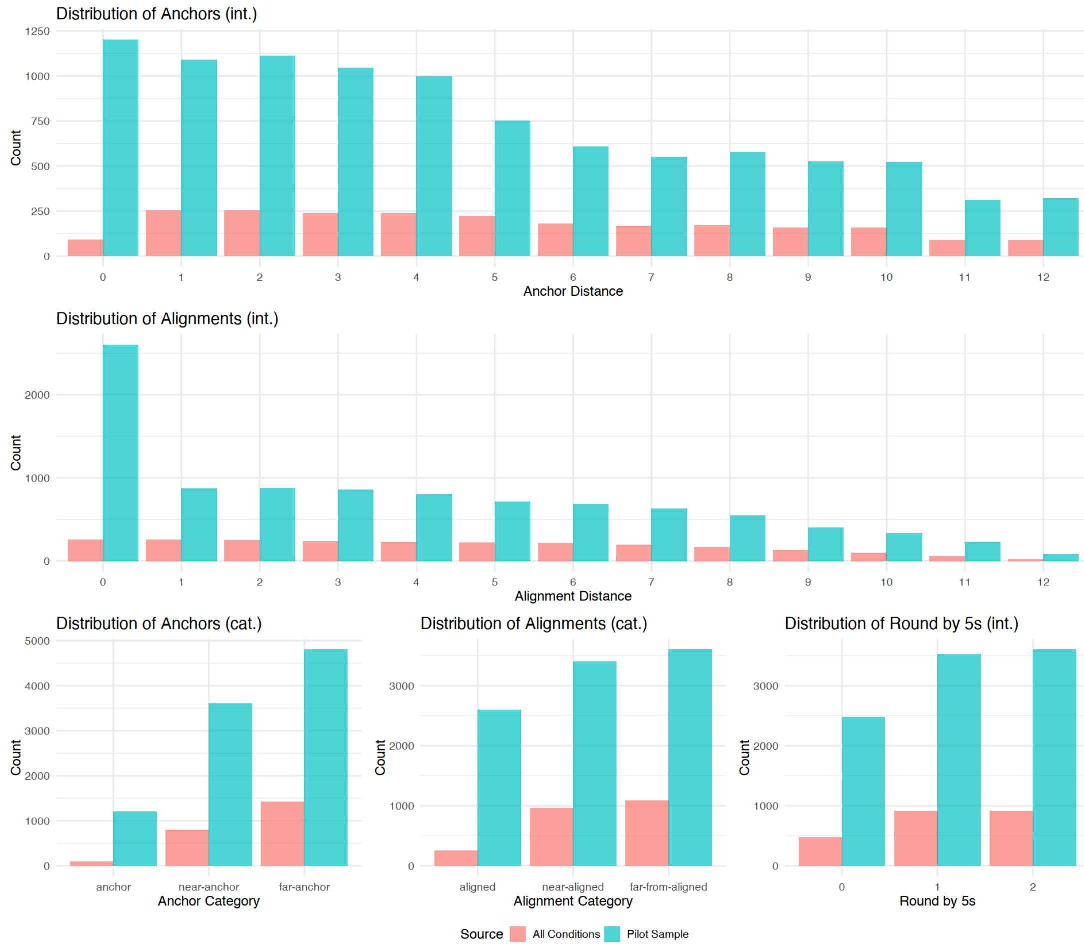


**Figure 1.** Distribution of counts of stimuli among all participants compared with the distribution of the total possible number of conditions of values and positions between 15 and 50 in the charts.

## 2 PROCEDURE EXHIBITS

the experiment consisted of the following screens: verification (Fig. 2), training instructions (Fig. 3), training trials (Fig. 4, Fig. 5), and experiment trials (Fig. 6, Fig. 7) Instructions given to participants:

Instructions

Your task is to estimate the marketshare of a company in the smartphone market to the nearest 1 percentage point using the chart provided. The data for each chart represents the total marketshare of smartphones for each of 7 companies during a particular time period. For simplicity, the companies are referred to as A, B, C, D, E, F, and G and make up 100% of the marketshare when combined.

Each question is a timed task. When you are ready to begin, please press Enter to view the question. Type your estimation in the input box and press Enter to submit your answer when you are confident in your estimation. Your goal is to provide a quick but accurate estimation, not to compute an exact answer.

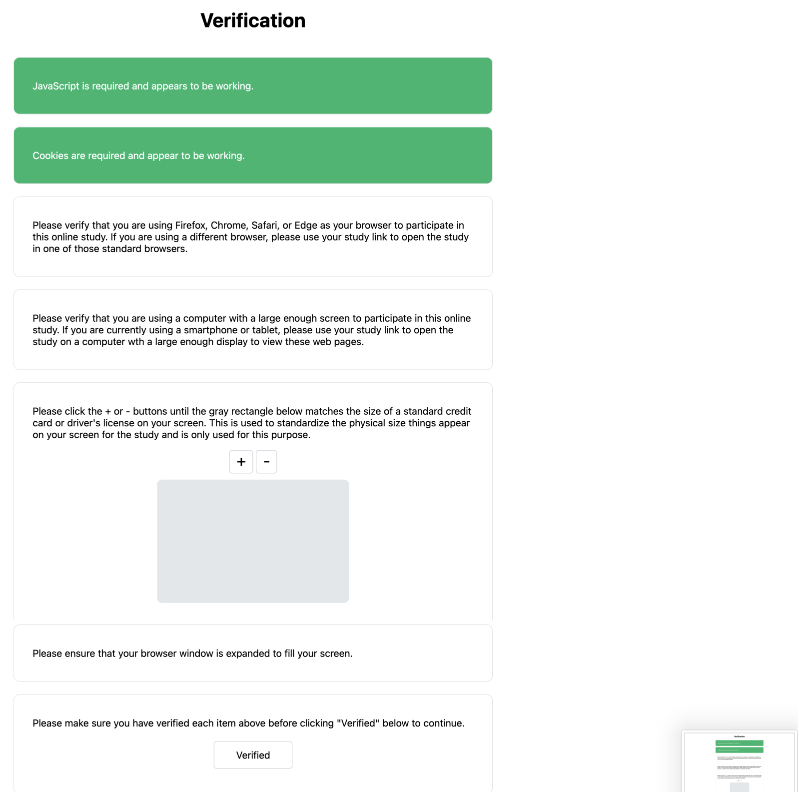


**Figure 2.** The verification page checks for JavaScript and browser cookies automatically. The participant resizes a rectangle to the size of a standard credit card or driver's license in order to standardize the size of the stimuli on each participants screen.

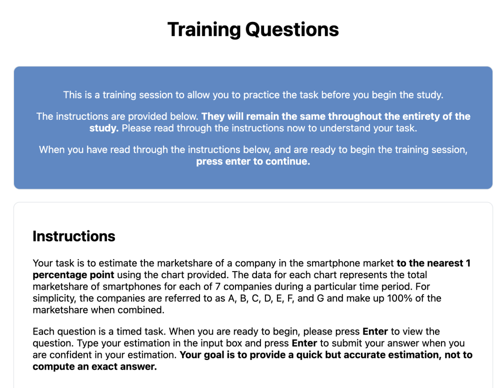


**Figure 3.** Training instructions screen.

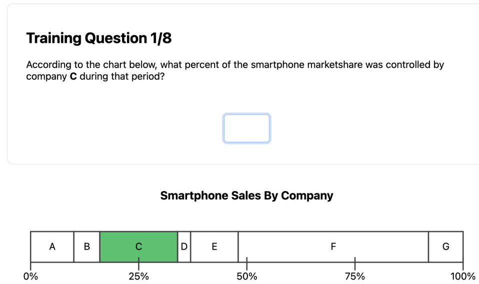


**Figure 4.** Training question example screen.

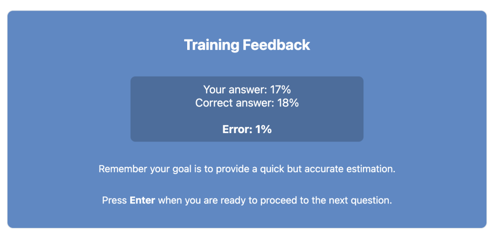


**Figure 5.** Training feedback example screen.

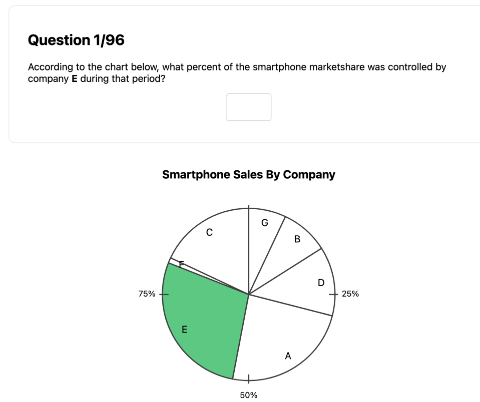


**Figure 6.** Example question with a pie chart screen.

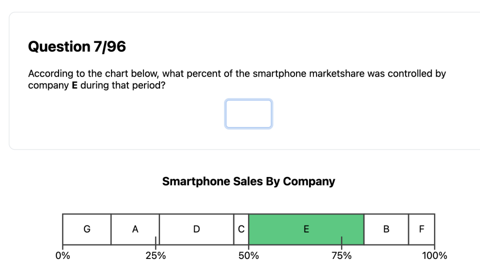


**Figure 7.** Example question with a line chart screen.

## 3 POWER ANALYSIS

In order to determine a sufficient number of participants to sample, and to preregister this criteria before running the study, we conducted a pilot study of ten participants and ran a power analysis on the results. In this analysis, we bootstrap sample using the pilot data to determine the point at which each effect reaches a power of 1 (a) and the sample size at which each effect becomes significant $p < 0.05$ (b). Based on this analysis, we decided to conduct the full experiment with 60 participants in order to have a sufficient sample size to effectively evaluate each of the measured effects.

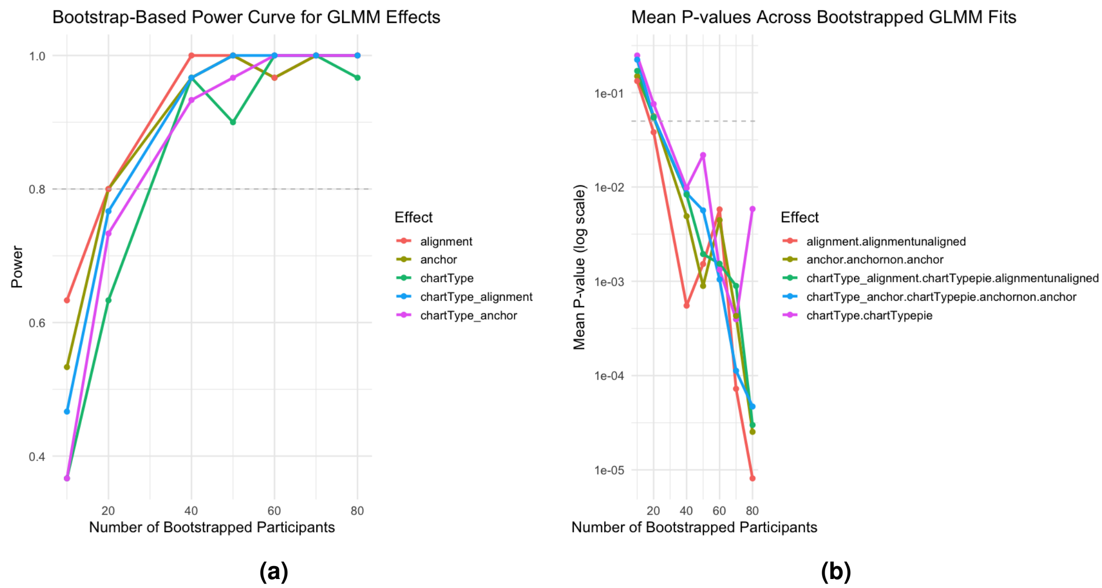


**Figure 8.** We conducted bootstrap sampling on pilot data of ten participants to determine an an appropriate sample size for our study. Figure a shows the power with the bootstrap number of participants and figure b shows the estimated p value for the number of participants. We found that 60 participants would provide sufficient power to detect a significant effect and conduct post hoc analyses.

# 4 EXCLUSIONS

We preregistered exclusion criteria for the study in order to remove outliers. The goal was to remove participants who provided answers that were unlikely to be meaningful (e.g., because of poor engagement or not understanding the instructions).

We determined to remove participants with mean absolute error outside of two median absolute deviations from the dataset as a whole (Fig. 9). 15 participants were replaced with participants who did not meet the exclusion criteria. Replacement participants were run and tested against the same criteria and replaced if they also failed to meet the preregistered standard. No additional participants were excluded for having more than 10% of responses falling outside of 2 standard deviations from the group as a whole.

Similarly, we excluded participants whose mean response time was outside of two median absolute deviations from the whole. 8 participants were excluded for this criteria(Fig. 10). No additional participants were excluded for having more than 10% of responses falling outside of 2 standard deviations from the group as a whole. One participant provided responses that were outside of the median absolute deviation for both time and error. Overall, out of the 60 original participants, 22 (36%) were replaced for meeting our preregistered exclusion criteria.


**Figure 9.** Of the 60 participants, 15 were excluded for providing responses with a mean absolute error outside of two median absolute deviations of the dataset as a whole. Top: Absolute error per-participant including the excluded participants. Means outside two MADs are highlighted in red. Responses outside of two standard deviations are highlighted in light red. Middle: Absolute error per-participant after exclusions are made. Bottom: Final absolute error results after exclusions. Outliers outside of two standard deviations are dropped from further analysis.


**Figure 10.** Of the 60 participants, 8 were excluded for providing responses with a mean response time outside of two median absolute deviations of the dataset as a whole. Top: Response time per-participant including the excluded participants. Means outside two MADs are highlighted in red. Responses outside of two standard deviations are highlighted in light red. Middle: Response time per-participant after exclusions are made. Bottom: Final response time results after exclusions. Outliers outside of two standard deviations are dropped from further analysis.

# 5 MODELS

## 5.1 Selected Models Fit

We fit GLMMs from the GLMMTMB R package for absolute error and response time. We provide the full model results here for absolute error and response time. The supplementary repository contains notebooks showing the code and outputs for all of the models: `https://github.com/uwgraphics/PartToWhole`.

### Absolute Error

$$|\text{error}| \sim \text{chartType} \times \text{alignment} \times \text{anchoring} + \ (1 + \text{alignment} + \text{anchoring} + \text{chartType} \mid \text{subject})$$

| Metric | AIC | BIC | logLik | -2*log(L) | df.resid |
|---|---|---|---|---|---|
| **Value** | 23202.9 | 23328.9 | -11582.5 | 23164.9 | 5593 |

**Table 1.** Model Fit Statistics

| Groups | Name | Variance | Std. Dev. | Corr | | |
|---|---|---|---|---|---|---|
| userID (Intercept) | | 0.49225 | 0.7016 | | | |
| | anchor | 0.13723 | 0.3704 | -0.92 | | |
| | alignment | 0.01488 | 0.1220 | -0.69 | 0.52 | |
| | chartType | 0.01849 | 0.1360 | 0.18 | -0.19 | -0.35 |

**Table 2.** Variance, Standard Deviation, and Correlations (Second Model)

| Term | Estimate | Std. Error | z value | $\mathbf{P}(>\|\mathbf{z}\|)$ |
|---|---|---|---|---|
| (Intercept) | 0.2668 | 0.0972 | 2.744 | 0.00607 ** |
| chartType | 0.0411 | 0.0345 | 1.193 | 0.23269 |
| anchor | 0.6437 | 0.0592 | 10.867 | $< 2 \times 10^{-16}$ *** |
| alignment | 0.3592 | 0.0344 | 10.437 | $< 2 \times 10^{-16}$ *** |
| chartType:anchor | 0.0117 | 0.0294 | 0.399 | 0.69009 |
| chartType:alignment | 0.0067 | 0.0290 | 0.232 | 0.81672 |
| anchor:alignment | -0.1714 | 0.0301 | -5.697 | $1.22 \times 10^{-8}$ *** |
| chartType:anchor:alignment | 0.0207 | 0.0289 | 0.717 | 0.47321 |

**Table 3.** Model Coefficients and Statistical Significance. ***: $p < 0.001$; **: $p < 0.01$.

#### Response Time

$$\text{time} \sim \text{chartType} \times \text{alignment} \times \text{anchoring} + \ (1 + \text{alignment} + \text{anchoring} + \text{chartType} \mid \text{subject})$$

| Metric | AIC | BIC | logLik | -2*log(L) | df.resid |
|---|---|---|---|---|---|
| **Value** | 27055.1 | 27181.1 | -13508.5 | 27017.1 | 5593 |

**Table 4.** Model Fit Statistics

| Groups | Name | Variance | Std. Dev. | Corr | | |
|---|---|---|---|---|---|---|
| userID (Intercept) | | 0.1248 | 0.3533 | | | |
| | anchor | 0.00199 | 0.04463 | 0.12 | | |
| | alignment | 0.00075 | 0.02733 | 0.31 | -0.44 | |
| | chartType | 0.00331 | 0.05755 | 0.16 | 0.35 | 0.04 |

**Table 5.** Variance, Standard Deviation, and Correlations

| Term | Estimate | Std. Error | z value | $P(>\lvert z \rvert)$ |
|---|---|---|---|---|
| (Intercept) | 1.8729 | 0.0464 | 40.41 | $< 2 \times 10^{-16}$ *** |
| chartType | 0.0009 | 0.0111 | 0.08 | 0.93788 |
| anchor | 0.1138 | 0.0101 | 11.31 | $< 2 \times 10^{-16}$ *** |
| alignment | 0.0447 | 0.0090 | 4.99 | $6.15 \times 10^{-7}$ *** |
| chartType:anchor | 0.0022 | 0.0082 | 0.27 | 0.78566 |
| chartType:alignment | 0.0076 | 0.0082 | 0.92 | 0.35589 |
| anchor:alignment | -0.0234 | 0.0082 | -2.84 | 0.00448 ** |
| chartType:anchor:alignment | -0.0107 | 0.0082 | -1.30 | 0.19339 |

**Table 6.** Model Coefficients and Statistical Significance. ***: $p < 0.001$; **: $p < 0.01$.

### 5.2 Evaluating Potential Covariates

Additionally, we compared these models to those including rounding to 5s and 10s as potential confounds of the absolute error measure:

#### Absolute Error

$$\begin{aligned} |\text{error}| \sim\ & \text{chartType} \times \text{alignment} \times \text{anchoring} + \text{roundBy5} \\ & + \ (1 + \text{alignment} + \text{anchoring} + \text{roundBy5} + \text{chartType} \mid \text{subject}) \end{aligned}$$

$$\begin{aligned} |\text{error}| \sim\ & \text{chartType} \times \text{alignment} \times \text{anchoring} + \text{roundBy10} \\ & + \ (1 + \text{alignment} + \text{anchoring} + \text{roundBy10} + \text{chartType} \mid \text{subject}) \end{aligned}$$

| Model | df | AIC | BIC | logLik | Deviance | Chisq | P(> Chisq) |
|---|---|---|---|---|---|---|---|
| model | 19 | 23202.92 | 23328.94 | -11582.46 | 23164.92 | NA | NA |
| modelRound5 | 25 | 23205.67 | 23371.49 | -11577.84 | 23155.67 | 9.25 | 0.1599 |
| modelRound10 | 25 | 23198.32 | 23364.13 | -11574.16 | 23148.32 | 7.35 | 0.0000 |

**Table 7.** ANOVA Comparison for models with no rounding variables, a distance to nearest 5 variable, and a distance to nearest 10 variable. We show that distance to nearest 5 does not provide statistically significant difference to the original model, but distance to nearest 10 does. Additionally, the model with distance to nearest 10 shows that factor to be statistically significant ($p < 0.05$)

### 5.3 Model Distribution Selections

We compared the Poisson, binomial, and Tweedie distributions for modeling the absolute error as it is a discrete, nonnegative, non-normal distribution:

| Model | df | AIC | BIC | logLik | Deviance | Chisq | P(> Chisq) |
|---|---|---|---|---|---|---|---|
| model_poisson | 18 | 26251.76 | 26371.15 | -13107.88 | 26215.76 | NA | NA |
| model_nb | 19 | 23202.92 | 23328.94 | -11582.46 | 23164.92 | 3050.841 | 0 |
| model_tweedie | 20 | 23580.43 | 23713.08 | -11770.22 | 23540.43 | 0.000 | 1 |

**Table 8.** Models using Poisson, Binomial, and Tweedie distributions were compared by an ANOVA Comparison for Models of their AIC statistics showing that the binomial model vastly outperforms the other models for fitting this data.

## 6 DEMOGRAPHICS

Participant demographic information was collected through Prolific and included information about participants, age, language, sex, ethnicity, country of birth, nationality, they're status as a student, and whether they were employed (see Fig. 11). The participants were constrained to be in the United States, fluent in English, have no colorblindness, and to be between 18 and 65 years of age in order to ensure ability to participate in understanding the instructions and completing the study.

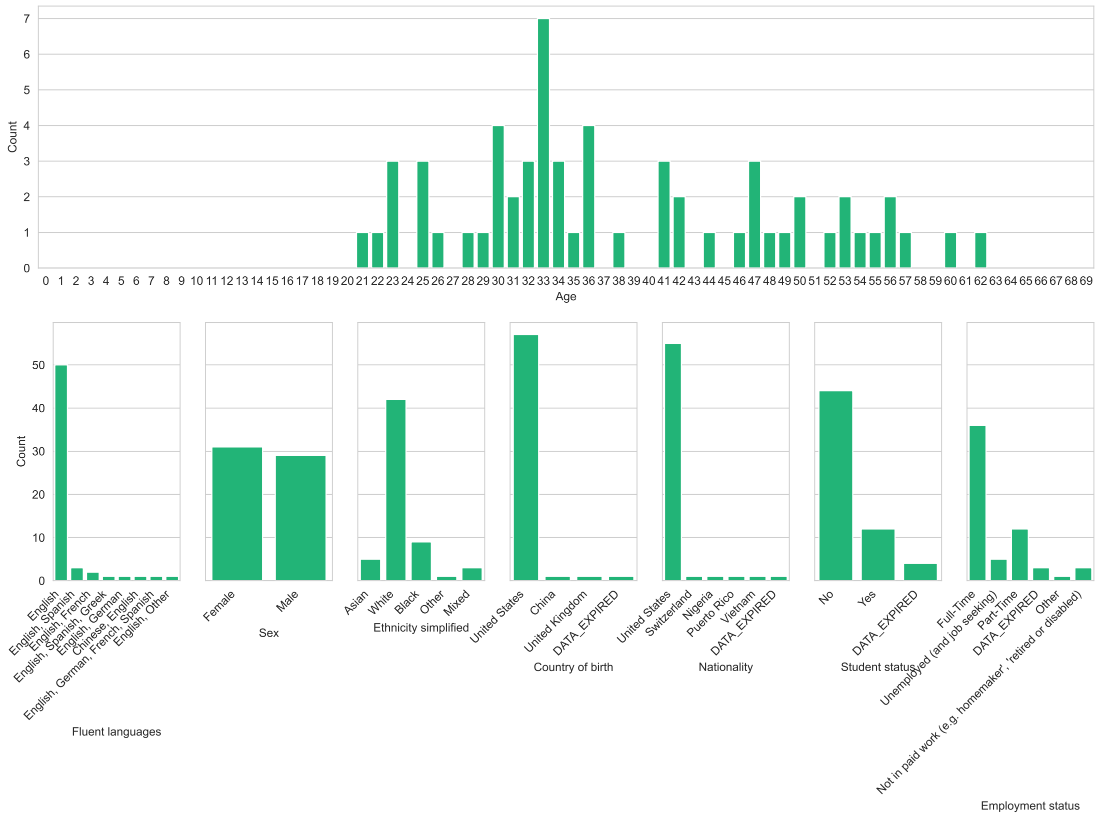


**Figure 11.** Distribution of participants across demographic measures provided by Prolific.