# Addendum to Pre-Analysis Plan
# Gender Discrimination Elicited in the General Population of the United States

Ingvild Almås Serena Cocciolo Jonathan de Quidt Sebastian Fest
Anna Sandberg

December 18, 2020

## 1 Background

In June 2020 we worked with the survey provider Gallup to collect a representative sample of the U.S population to act as employers for our experiments. However, due to miscommunication with Gallup, data collection was carried out by using a simple random sampling procedure rather than a stratified random sampling procedure, as was intended originally. Due to variation in response rates, the raw data we received turned out to be far from representative of the US population. Specifically, in order to achieve representativeness of our sample, a heavy re-weighting, using poststratification weights, was required. The outcome of this reweighting procedure gave rise to the problem that a small number of observations in the collected sample had a disproportionately large influence.

In order to illustrate the problem, we present two indicators of the degree of reweighting required to adjust the data. Figure 1 is a density plot of the age distribution in the June 2020 sample, alongside a density plot based on US Census data. It is apparent that the June 2020 sample skews significantly older than the population.

The second indicator is the distribution of poststratification weights required to restore representativeness in this sample, plotted in Figure 2. Weights are constructed such that the mean weight is 1. Approximately 60% of the sample are bunched together with very low weights at an approximate weight of 0.16. In contrast, 7% of the sample receive very high weights at about 6.32. The remainder of observations thinly spread in between. This pattern of weights implies that the 7% of the sample with the highest weight end up receiving 45% of the overall weight in the sample, while the 60% of the sample with the lowest weight end up receiving less than 10% of the overall weight in the sample.[1]

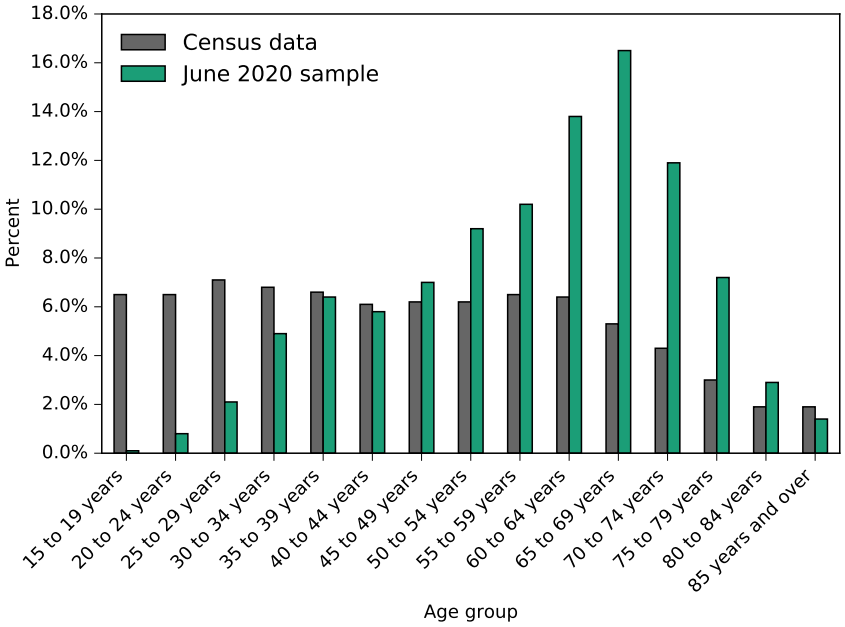
Although reweighting can in principle restore representativeness, any results will be highly imprecise and sensitive to those 7% of observations and how they are distributed across our 9 treatments. Moreover we were not comfortable with putting high certainty on results that were so heavily dependent on poststratification weights, since these are constructed ex-post after observing patterns of selection into the study.

## 2 Addendum to Plan

We have had a very good and constructive collaboration with the Gallup team and when realizing the issues with the skewness of the post-stratification weights, they offered to replace the June 2020

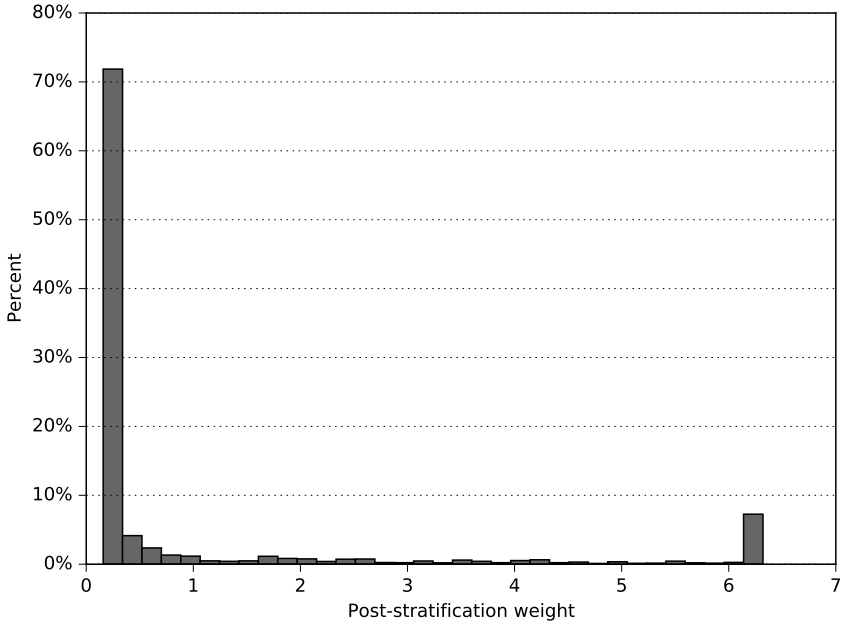
[1] 321 observations have the highest weight of 6.319878 and 2,732 observations have the lowest weight of 0.157997. The weights are computed such that the sum of all weights equals the total sample size of 4,534. (321*6.319878)/4,534=0.447434. (2,732*0.157997)/4,534=0.095202.

Figure 1: Age distribution.




*Note:* The figure plots the distribution of age for subjects from the U.S population in 2019 and the sample data that we obtained from Gallup for the experiment.

Figure 2: Distribution of post-stratification weights.




*Note:* The figure plots the distribution of post-stratification weights in the sample data that we obtained from Gallup for the experiment.

sample with an entirely new sample. This new data collection is scheduled to take place in January 2021 and will target 4,450 employer participants as before. The sampling will be pre-stratified to be representative on age, gender, race/ethnicity, education, and census region.

To adjust for any remaining nonrepresentativeness, Gallup will provide us with prestratification and poststratification weights corresponding to this new sample. Due to the stratified sampling strategy, we expect the influence of these weights to be relatively minimal, unlike in the June 2020 sample.

**The substantive pre-committed decision in this addendum is that all of our main analysis in the research papers will use the January 2021 sample only.**

The rationale for this choice is twofold. First, any attempt to combine the samples needs to take a stand on how to weight them, and there is no obvious way to do this that doesn't end up either heavily weighting the 7% of observations with large weights in June 2020 (e.g. if we weight the samples equally, those observations will receive 22.5% of the overall weight), or heavily weighting the January 2020 sample. It is undesirable for our main analysis to rely on an arbitrary choice of weighting strategy. Second, a lot has happened between June 2020 and January 2021 (the pandemic dynamics have evolved, and the United States has conducted a Presidential election). Focusing on the January 2021 sample will tell us about behavior and treatment effects conditional on the events that have passed. Trying to combine the samples might lead our results to depend on possible changes in attitudes or treatment effects over that period. How any such changes enter the main findings will be dependent on how we weight the two periods.

For completeness, in Appendix material we will supply robustness checks using the combined (June 2020 and January 2021) samples. Gallup will provide us with poststratification weights that render this combined sample representative of the US population. These weights can still be expected to be skewed, because we will be mixing a stratified sample with a nonstratified sample, but we expect less skewness than in the June 2020 sample alone.