Irregularities in LaCour (2014)

David Broockman, Assistant Professor, Stanford GSB (as of July 1),
dbroockman@stanford.edu
Joshua Kalla, Graduate Student, UC Berkeley, kalla@berkeley.edu
Peter Aronow, Assistant Professor, Yale University, peter.aronow@yale.edu

May 19, 2015

Summary

We report a number of irregularities in the replication dataset posted for LaCour and Green (*Science*, "When contact changes minds: An experiment on transmission of support for gay equality," 2014) that jointly suggest the dataset (LaCour 2014) was not collected as described. These irregularities include baseline outcome data that is statistically indistinguishable from a national survey and over-time changes that are unusually small and indistinguishable from perfectly normally distributed noise. Other elements of the dataset are inconsistent with patterns typical in randomized experiments and survey responses and/or inconsistent with the claimed design of the study. A straightforward procedure may generate these anomalies nearly exactly: for both studies reported in the paper, a random sample of the 2012 Cooperative Campaign Analysis Project (CCAP) form the baseline data and normally distributed noise are added to simulate follow-up waves.

Timeline of Disclosure

- January - April, 2015. Broockman and Kalla were impressed by LaCour and Green (2014) and wanted to extend the article's methodological and substantive discoveries. We began to plan an extension. We sought to form our priors about several design parameters based on the patterns in the original data on which the paper was based, LaCour (2014). As we examined the study's data in planning our own studies, two features surprised us: voters' survey responses exhibit much higher test-retest reliabilities than we have observed in any other panel survey data, and the response and reinterview rates of the panel survey were significantly higher than we expected. We set aside our doubts about the study and awaited the launch of our pilot extension to see if we could manage the same parameters. LaCour and Green were both responsive to requests for advice about design details when queried.
- May 6, 2015. Broockman and Kalla launch a pilot of the extension study.
- May 15, 2015. Our initial questions about the dataset arose as follows. The response rate of the pilot study was notably lower than what LaCour and Green (2014) reported. Hoping we could harness the same procedures that produced the original study's high reported response rate, we attempt to contact the survey firm we believed had performed the original study and ask to speak to the staffer at the firm who we believed helped perform Study 1 in LaCour and Green (2014). The survey firm claimed they had no familiarity with the project and that they had never had an employee with the name of the staffer we were asking for. The firm also denied having the capabilities to perform many aspects of the recruitment procedures described in LaCour and Green (2014).
- May 15, 2015. Broockman and Kalla return to the data and uncover irregularities 3, 4, 5, and 6 below and describe the findings to Green. Green expresses concern and suggests several avenues of further investigation, one of which led to the discovery of irregularity 7.
- May 16, 2015. To ensure we were correctly implementing one of Green's suggestions, Broockman and Kalla ask Aronow to help to confirm and expand the data analysis. (Aronow has statistical expertise in the field and has coauthored a working paper that included data from LaCour (2014).)
- May 16, 2015. Broockman suspects the CCAP data may form the source distribution and Kalla finds the CCAP data in the AJPS replication archive for an unrelated paper. Irregularities 1 and 8 emerge.

- May 16, 2015. Broockman, Kalla, and Aronow disclose irregularities 1, 7, and 8 to Green. Green requests this report and the associated replication files.
- May 16, 2015. Aronow discovers irregularity 2.
- May 17, 2015. Broockman, Kalla, and Aronow prepare and send this report and replication code to Green. Green reads and agrees a retraction is in order unless LaCour provides countervailing evidence. Green also requests this report be made public concurrently with his retraction request, if this request is deemed appropriate.
- May 18-9, 2015. Green conveys to Aronow and Broockman that LaCour has been confronted and has confessed to falsely describing at least some of the details of the data collection. The authors of this report are not familiar with the details of these events.
- May 19, 2015. Green posts a public retraction of LaCour and Green (2014) on his website.
- May 19, 2015. Green submits a retraction letter to *Science*. Green sends us the retraction letter and asks that we post the report afterwards. He agrees the retraction letter can be part of the report. We post this report. The retraction letter is included as an Appendix.
- May 19, 2015. The replication data no longer appears available at https://www.openicpsr.org/repoEntity/show/24342. A screenshot of the page when the data was available on May 18 is available in the replciation files for this report.

Background on LaCour (2014) and LaCour and Green (2014)

LaCour and Green (2014, *Science*) report a remarkable result: a ~20-minute conversation with a gay canvasser produces large positive shifts in feelings towards gay people that persist for over a year. The study's design is also notable: over 12% of voters invited to participate in the ostensibly unrelated survey that formed the study's measurement apparatus agreed to be surveyed; nearly 90% were successfully reinterviewed; and each voter referred an average of 1.33 other voters to be part of the study who lived in the study area. The paper is based on a dataset that allegedly describes two field experiments, LaCour (2014).

Data

We downloaded LaCour (2014) from https://www.openicpsr.org/repoEntity/show/24342 on May 15, 2015. Our copy of this dataset as well as the source code for this report is available at https://web.stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities_replication_files.zip.

The data allegedly consist of repeated observations of the same 11,948 voters over a series of weeks. Two dependent variables were captured at each wave: a 5-point policy preference item on same-sex marriage (SSM) and a 101-point feeling thermometer towards gay men and lesbians.

Summary of Irregularities

No one of the irregularities we report alone constitutes definitive evidence that the data were not collected as described. However, the accumulation of many such irregularities, together with a clear alternative explanation that fits the data well, leads us to be skeptical the data were collected as described.

1. The article claims that both studies were drawn from two distinct, non-random, snowball samples of voters in Los Angeles County, California. However, the distribution of the gay feeling thermometer in both studies is identical to the same feeling thermometer in a national survey dataset to which the author had access. However, it differs strongly from a variety of reference distributions of this item from other datasets.

2. The joint distributions of the feeling thermometer and the same-sex marriage policy item are identical in the two studies despite the fact that they are allegedly drawn from two distinct, non-random, snowball samples.
3. The feeling thermometer is a notoriously unreliable instrument, showing a great deal of measurement error. However, in both studies, respondents' feeling thermometer values are extremely reliable – more so than nearly any other survey items of which we are aware.
4. The changes in respondents' feeling thermometer scores are perfectly normally distributed. Not one respondent out of thousands provided a response that meaningfully deviated from this distribution.
5. In general, feeling thermometer responses show predictable heaping patterns, such as a concentration of responses at 50. These patterns are present in the first wave but none of the follow-up waves in the data; the heaped responses in the first wave gain the same normally distributed noise as described above.
6. The point above could be explained by changes in item format between the first and follow-up waves, but all of the follow-up waves differ from the baseline wave by independent normal distributions; that is, differences between wave 1 and each follow-up wave do not manifest in subsequent follow-up waves. This is not consistent with changes in item format, which should generate non-random measurement error that would lead to correlations between the items in the new format. It is consistent with each wave being simulated as the first wave plus an independent normal distribution.
7. The voters that the dataset indicates the campaign successfully contacted have identical attitudes to the voters in the treatment group the campaign did not successfully contact; usually in experiments, voters who can be successfully reached differ in systematic ways.
8. All the above patterns can be explained by an extremely simple data generating process with the 2012 Cooperative Campaign Analysis Project (CCAP) data as its starting point.

Below we provide replication code that reveals these irregularities.

Replication of main result of Study 2

It remains possible that the wrong replication data were posted; for example, perhaps simulated data were generated for a course. To help evaluate whether the data used in the paper were posted, we first replicate the main result of Study 2 to verify that we have loaded and processed the data correctly. The below replicates the point estimate of 6.8 from the Note in the bottom of Table S13 exactly.

```r
set.seed(63624) # from random.org
lacour <- read.csv('24343.csv')
# https://www.openicpsr.org/repoEntity/show/24342
lacour.therm <- subset(lacour, wave == 1)$Therm_Level
lacour.reshaped <- reshape(lacour, v.names = c('Therm_Level',
      'Therm_Change', 'SSM_Level', 'SSM_Change'), idvar =
      'panelist_id', timevar = 'wave', direction = 'wide')
summary(lm(Therm_Level.2 ~ Therm_Level.1 + (TA == "G by G"),
           data = subset(lacour.reshaped, STUDY == "Study 2")))
```

```
##
## Call:
## lm(formula = Therm_Level.2 ~ Therm_Level.1 + (TA == "G by G"),
##     data = subset(lacour.reshaped, STUDY == "Study 2"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.194  -4.476  -0.476   3.688  35.558
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.727135   0.423275    4.08 4.66e-05 ***
## Therm_Level.1       0.965808   0.005973  161.69  < 2e-16 ***
## TA == "G by G"TRUE 2.458857   0.341052    7.21 7.76e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.87 on 2129 degrees of freedom
##   (309 observations deleted due to missingness)
## Multiple R-squared:  0.925,  Adjusted R-squared:  0.9249
## F-statistic: 1.313e+04 on 2 and 2129 DF,  p-value: < 2.2e-16
```

```
lacour.iv <- subset(lacour.reshaped, STUDY == "Study 2" & !is.na(Therm_Change.2))
lacour.iv$D <- lacour.iv$Contact!="None"
lacour.iv$Z <- lacour.iv$TA == "G by G"
summary(ivreg(Therm_Change.2 ~ D | Z, data = lacour.iv))
```

```
##
## Call:
## ivreg(formula = Therm_Change.2 ~ D | Z, data = lacour.iv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.5463  -4.5967   0.2522   4.2522  31.4537
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2522     0.2354  -1.071    0.284
## DTRUE         6.7985     0.9237   7.360 2.61e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.587 on 2130 degrees of freedom
## Multiple R-Squared: 0.105,   Adjusted R-squared: 0.1045
## Wald test: 54.17 on 1 and 2130 DF,  p-value: 2.614e-13
```

1. Similarity with 2012 CCAP Data

We will begin with what we see as the most compelling evidence that raises questions that the data were collected as described: the distribution of one of the study's main dependent variables is identical to the distribution of that variable in the 2012 Cooperative Campaign Analysis Project, a national survey.

Our suspicions that the data might have been lifted from CCAP arose as follows. Given that the baseline distribution looked like genuine data in terms of heaping patterns, we suspected it might have come from some other real source. We suspected this source might be CCAP because a) LaCour and Green (2014) notes that the wording for the question was taken from the CCAP, b) the paper describes a result from CCAP data, so we knew the authors had access to the dataset, c) the CCAP dataset is not widely available, making it a potentially appealing source, and d) the CCAP is one of few datasets with a feeling thermometer question about gay men and lesbians.

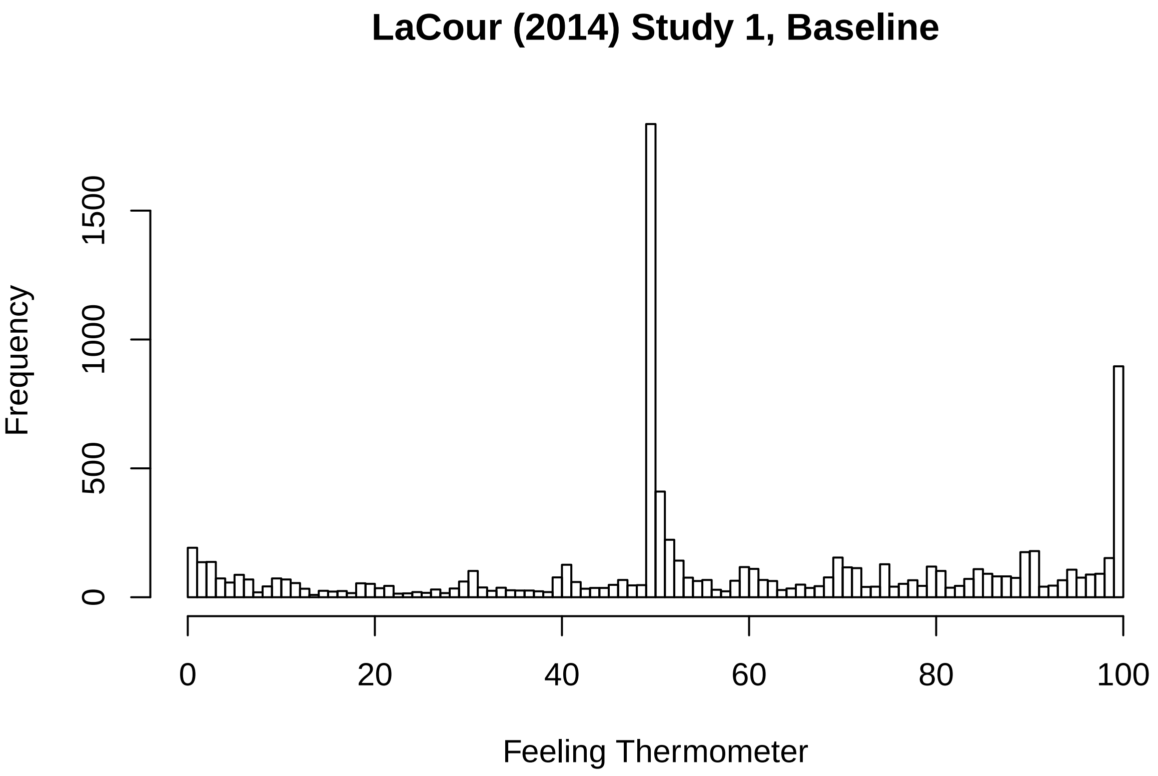
The CCAP dataset is not generally publicly available. We gained access to it because an unrelated article posted the dataset in its replication files (Tesler 2014; see ccap12short.tab at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26721).

The copy of CCAP posted at that site does not have the 5-point same-sex marriage question, so our analysis is restricted to the 101-point feeling thermometer.

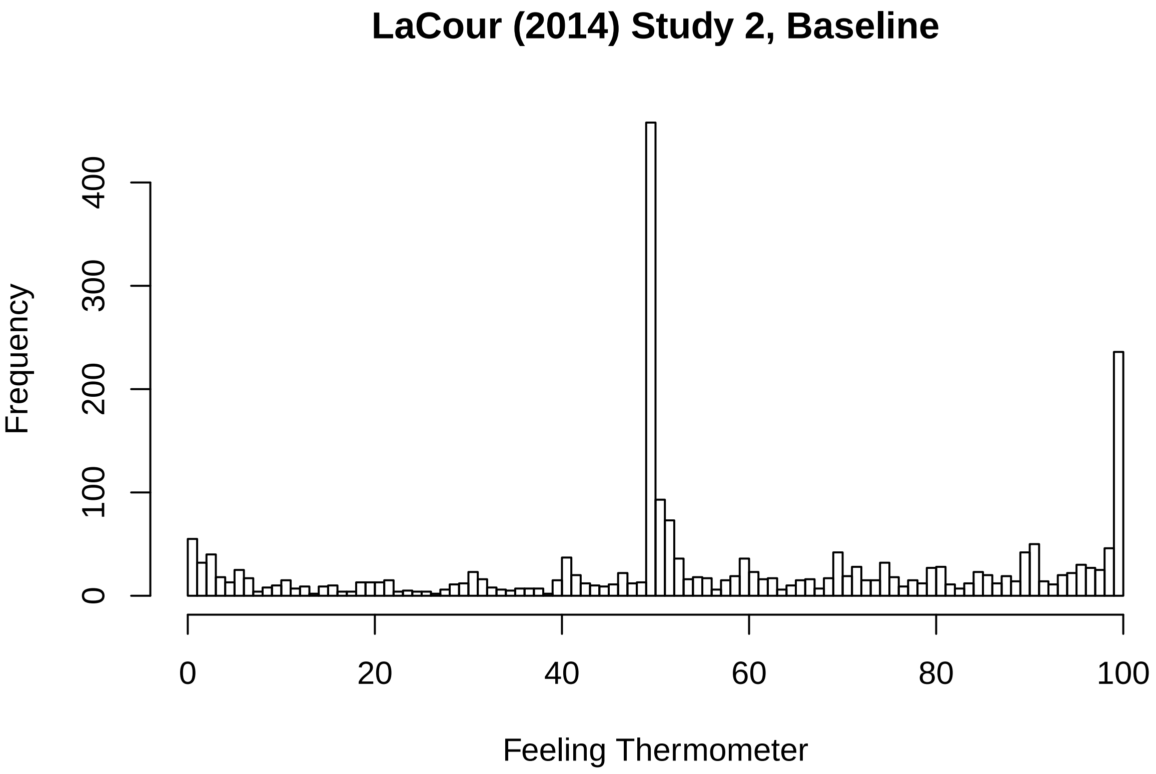
There were many NAs in the CCAP dataset and none in the LaCour (2014) dataset. We were unable to locate discussion of how no answers were dealt with in LaCour and Green (2014) or LaCour (2015). We recoded the NAs in the CCAP dataset to values of 50, as this is the general convention and where feeling thermometers typically begin by default on web surveys.

Below is the distribution of the feeling thermometer in the CCAP and in the baseline wave of LaCour (2014), split out by study. The paper claims that, for Study 2, “a new subject pool of panel respondents was recruited in a different region of Los Angeles County using the same criteria as in Study 1.” Los Angeles County is diverse and it would be highly surprising if two distinct, non-random samples were to be statistically identical. However, we find that Study 1’s and Study 2’s respondents had the exact same distribution of responses to the feeling thermometer as each other and as the CCAP respondents.

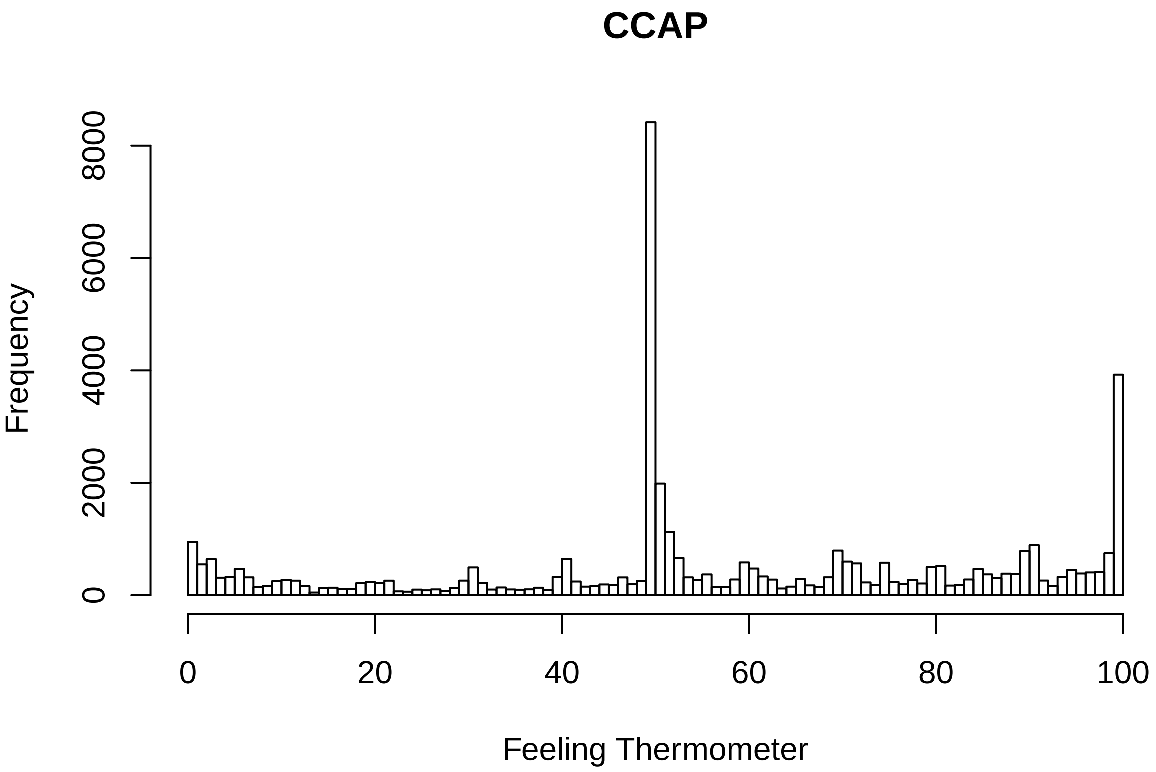
```r
ccap.data <- read.table('ccap/ccap12short.tab', sep="\t", header=TRUE)
ccap.therm <- ccap.data$gaytherm
ccap.therm[is.na(ccap.therm)] <- 50

lacour.therm.study1 <- subset(lacour, wave == 1 & STUDY == "Study 1")$Therm_Level
lacour.therm.study2 <- subset(lacour, wave == 1 & STUDY == "Study 2")$Therm_Level

hist(lacour.therm.study1, breaks=101, xlab="Feeling Thermometer",
     main = "LaCour (2014) Study 1, Baseline")
```



```
hist(lacour.therm.study2, breaks=101, xlab="Feeling Thermometer",
     main = "LaCour (2014) Study 2, Baseline")
```



```
hist(ccap.therm, breaks=101, main="CCAP", xlab="Feeling Thermometer")
```

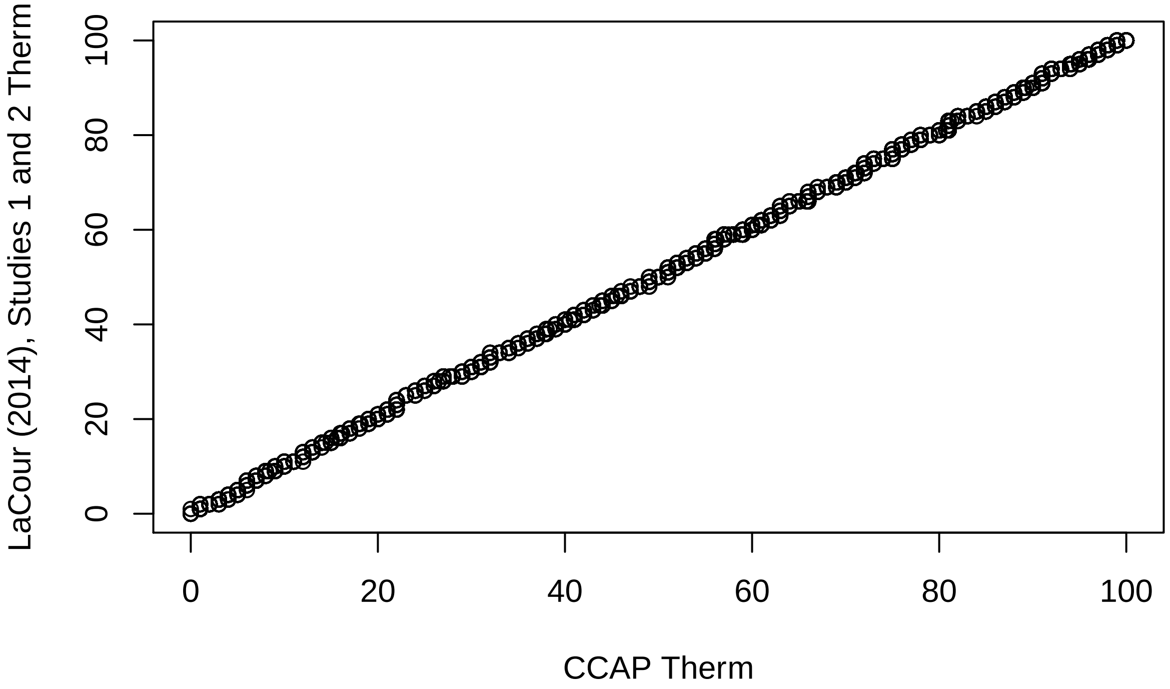


The distributions not only look very similar, they are statistically indistinguishable. A Kolmogorov-Smirnov test finds no differences between LaCour (2014) and the CCAP data, and plotting the quantiles of the two data sources against each other yields a strikingly uniform pattern.

```
ks.test(lacour.therm, ccap.therm)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  lacour.therm and ccap.therm
## D = 0.0087, p-value = 0.4776
## alternative hypothesis: two-sided
```

```
qqplot(ccap.therm, lacour.therm, ylab = "LaCour (2014), Studies 1 and 2 Therm", xlab = "CCAP Therm")
```

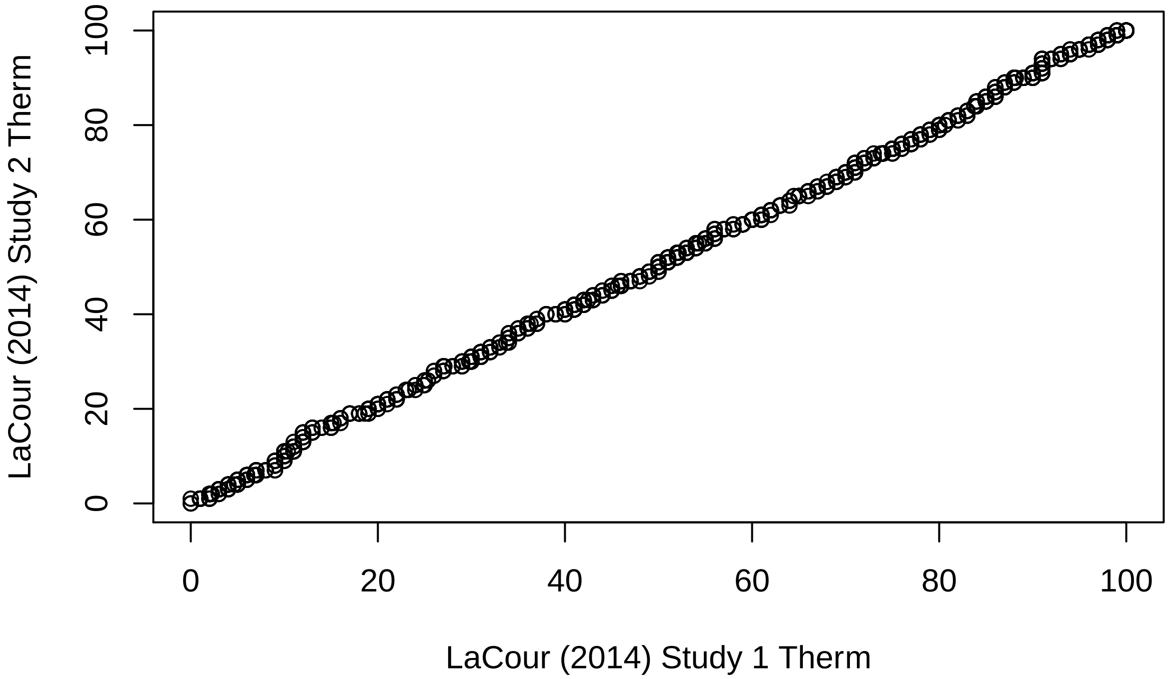


The two studies in the paper also have indistinguishable baseline values despite having been allegedly drawn from different non-random samples.

```
ks.test(lacour.therm.study1, lacour.therm.study2)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  lacour.therm.study1 and lacour.therm.study2
## D = 0.0139, p-value = 0.8458
## alternative hypothesis: two-sided
```

```
qqplot(lacour.therm.study1, lacour.therm.study2, xlab = "LaCour (2014) Study 1 Therm",
       ylab = "LaCour (2014) Study 2 Therm")
```



Difference Between LaCour (2014) and Reference Distributions

One way in which the data might exactly match the national marginals would be if thermometer response data were stable across contexts and subject pools. To assess this possibility, we compare the distribution in the study with six other reference distributions of this item, two from surveys with non-random sampling we have conducted in Philadelphia and Miami (which employed the same question wording and that we analyze using two different rules for recoding non-answers), three from the American National Election Study national sample, and the final by subsetting the CCAP data to just the California CCAP sample. All of these reference distributions yield KS tests and QQ-plots that are markedly different. In light of these differences we see it as unlikely that two non-random samples in Los Angeles County would yield such striking similarities, both to each other and to the national CCAP sample.

```
#Philadelphia, 2015
philly <- read.csv('other therms/philly_therm.csv')$gaytherm
philly.nas.recoded <- philly
philly.nas.recoded[is.na(philly.nas.recoded)] <- 50

#Miami, 2015
miami <- read.csv('other therms/miami_therm.csv')$gaytherm
miami.nas.recoded <- miami
miami.nas.recoded[is.na(miami.nas.recoded)] <- 50

#ANES, 2000-2002-2004 Panel Study
library(foreign)
anes <- read.dta('anes_mergedfile_2000to2004_dta/anes_mergedfile_2000to2004.dta')
anes2000 <- anes$M001321
anes2002 <- anes$M025067
anes2004 <- anes$M045035

ks.test(anes2000, lacour.therm)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  anes2000 and lacour.therm
## D = 0.2471, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
ks.test(anes2002, lacour.therm)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  anes2002 and lacour.therm
## D = 0.2781, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
ks.test(anes2004, lacour.therm)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  anes2004 and lacour.therm
## D = 0.255, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
ks.test(philly, lacour.therm)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  philly and lacour.therm
## D = 0.2757, p-value = 3.861e-08
## alternative hypothesis: two-sided
```

```r
ks.test(philly.nas.recoded, lacour.therm)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  philly.nas.recoded and lacour.therm
## D = 0.2106, p-value = 7.846e-06
## alternative hypothesis: two-sided
```

```r
ks.test(miami, lacour.therm)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  miami and lacour.therm
## D = 0.1656, p-value = 0.01079
## alternative hypothesis: two-sided
```

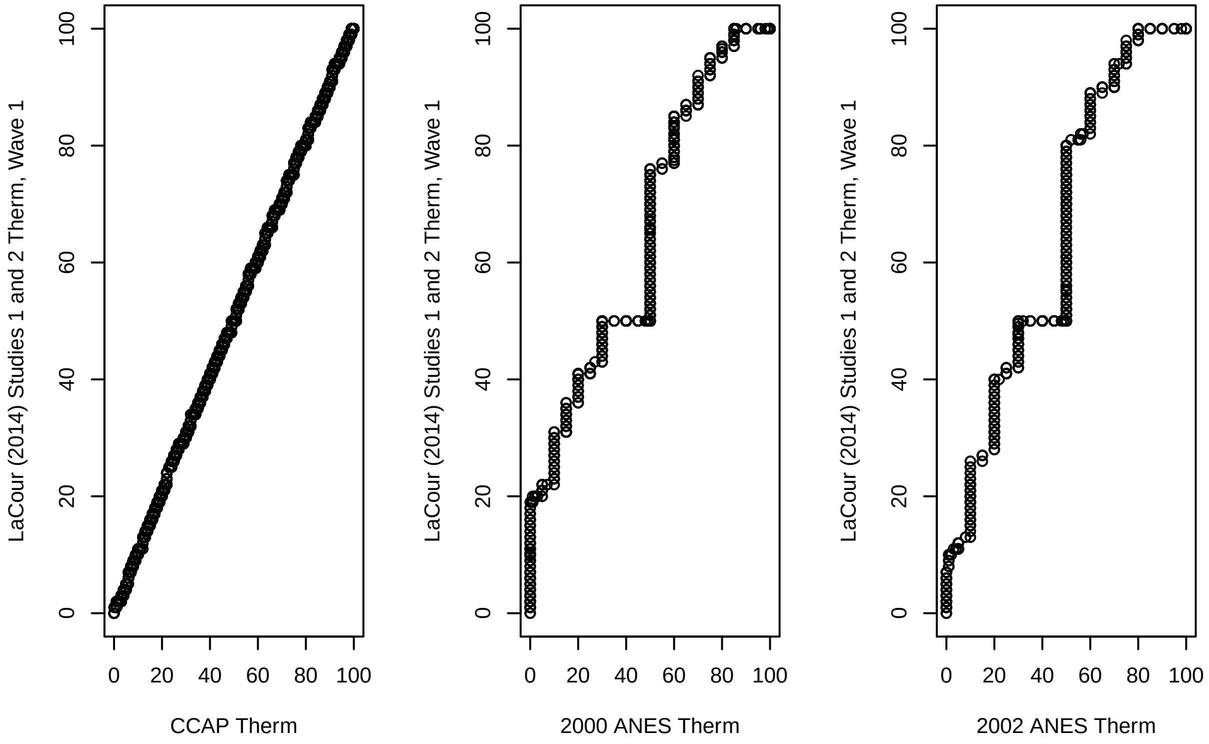
```r
ks.test(miami.nas.recoded, lacour.therm)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  miami.nas.recoded and lacour.therm
## D = 0.2452, p-value = 5.462e-07
## alternative hypothesis: two-sided
```

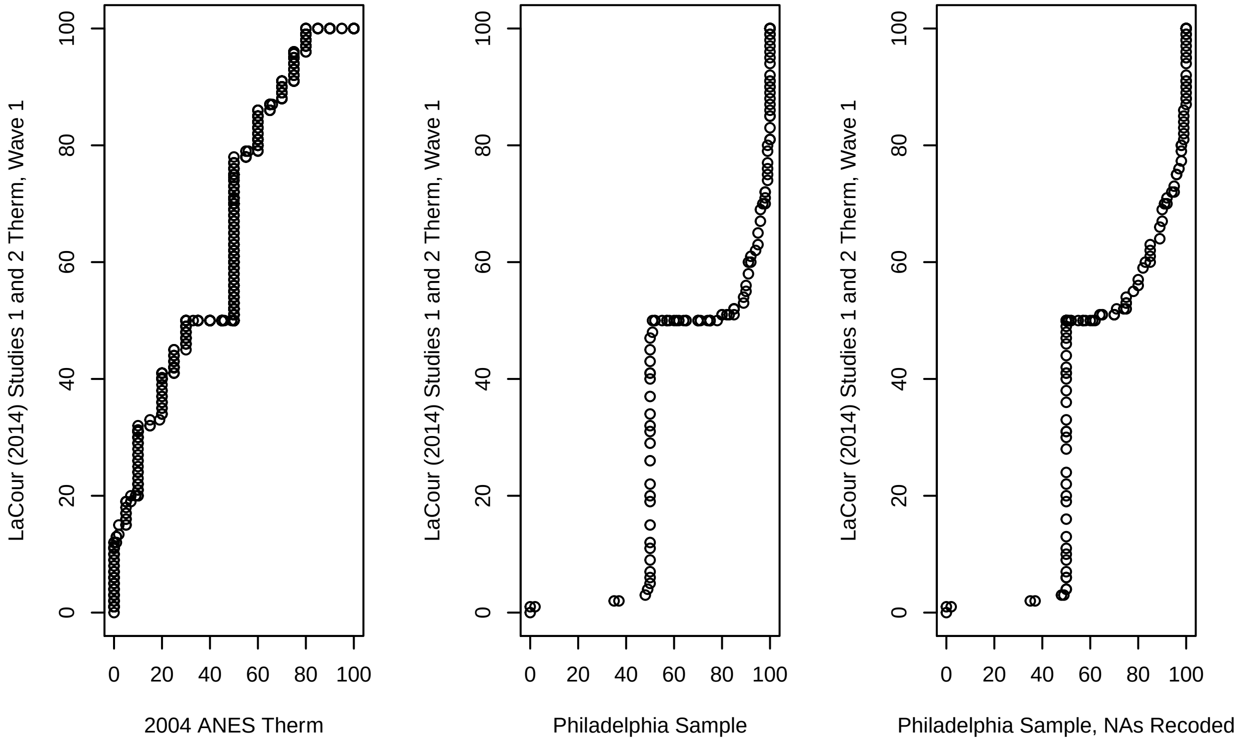
```r
## CCAP, California only
ccap.therm.ca <- ccap.therm[ccap.data$inputstate==6]
ks.test(ccap.therm.ca, lacour.therm)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ccap.therm.ca and lacour.therm
## D = 0.06, p-value = 1.843e-10
## alternative hypothesis: two-sided
```

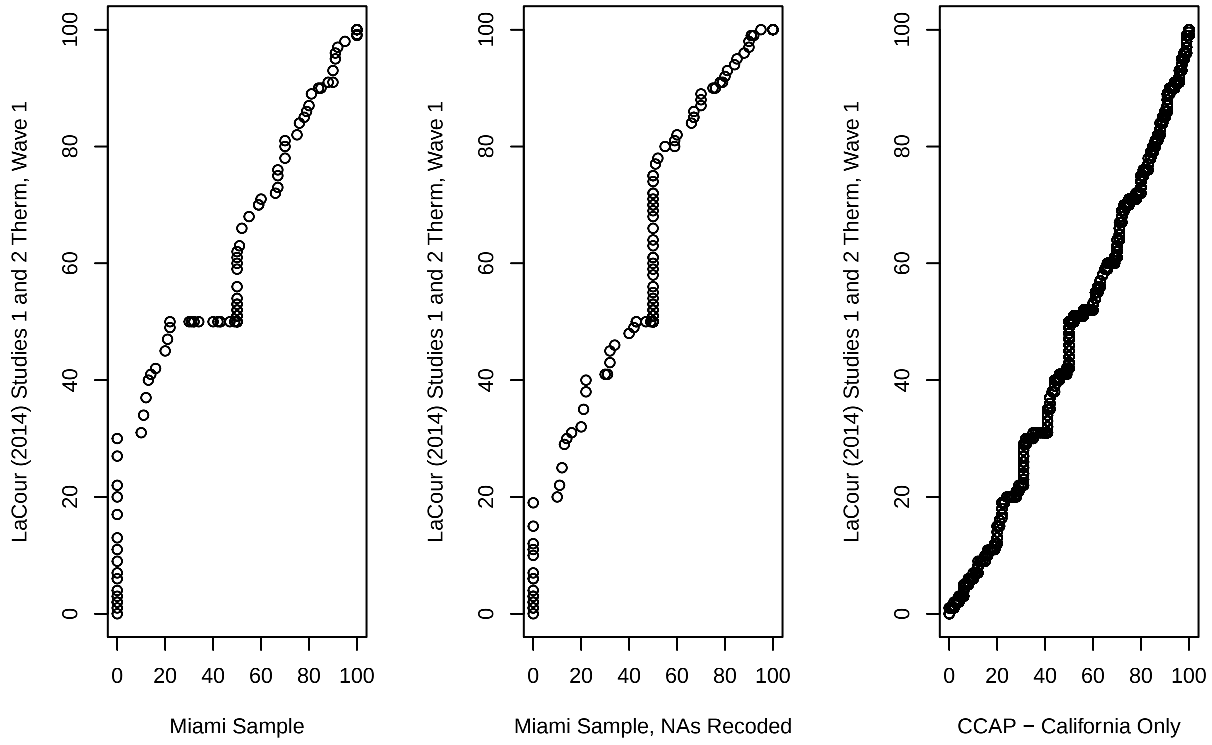
```r
par(mfrow=c(1,3))
qqplot(ccap.therm, lacour.therm, xlab = "CCAP Therm",
       ylab = "LaCour (2014) Studies 1 and 2 Therm, Wave 1")
qqplot(anes2000, lacour.therm, xlab = "2000 ANES Therm",
       ylab="LaCour (2014) Studies 1 and 2 Therm, Wave 1")
qqplot(anes2002, lacour.therm, xlab = "2002 ANES Therm",
       ylab="LaCour (2014) Studies 1 and 2 Therm, Wave 1")
```



```r
par(mfrow=c(1,3))
qqplot(anes2004, lacour.therm, xlab = "2004 ANES Therm",
       ylab="LaCour (2014) Studies 1 and 2 Therm, Wave 1")
qqplot(philly, lacour.therm, xlab = "Philadelphia Sample",
       ylab="LaCour (2014) Studies 1 and 2 Therm, Wave 1")
qqplot(philly.nas.recoded, lacour.therm, xlab = "Philadelphia Sample, NAs Recoded",
       ylab="LaCour (2014) Studies 1 and 2 Therm, Wave 1")
```



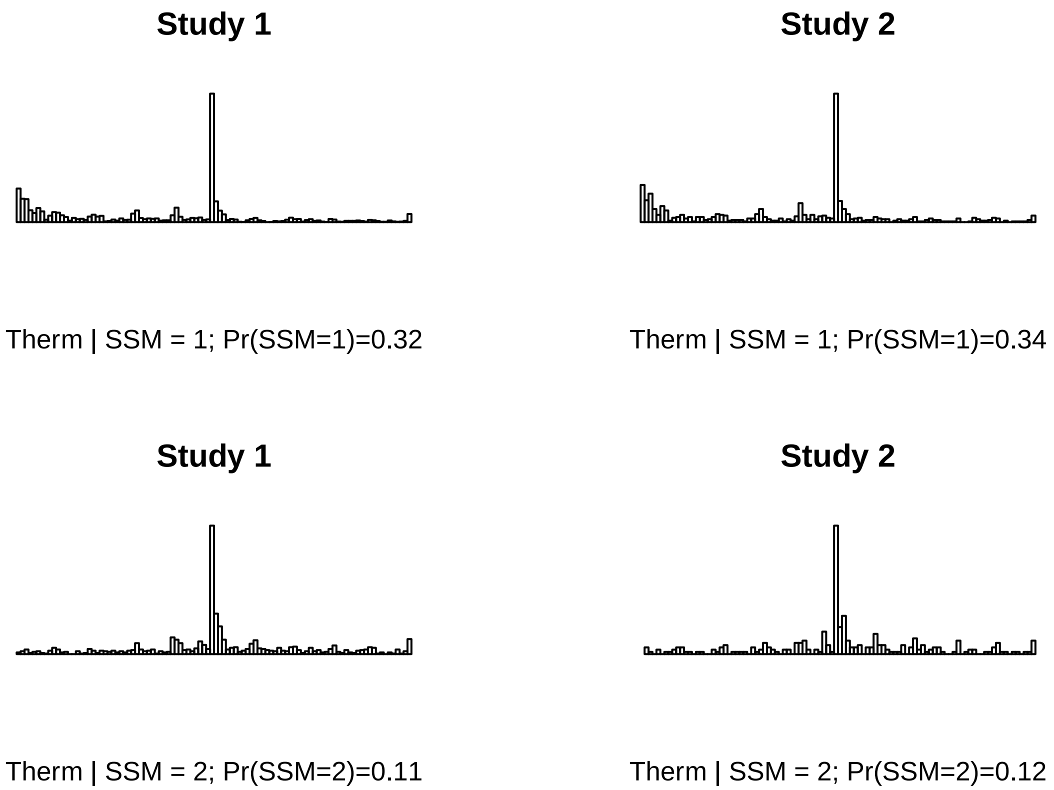
```r
par(mfrow=c(1,3))
qqplot(miami, lacour.therm , xlab = "Miami Sample", ylab="LaCour (2014) Studies 1 and 2 Therm, Wave 1")
qqplot(miami.nas.recoded, lacour.therm , xlab = "Miami Sample, NAs Recoded",
       ylab="LaCour (2014) Studies 1 and 2 Therm, Wave 1")
qqplot(ccap.therm.ca, lacour.therm, xlab = "CCAP - California Only",
       ylab = "LaCour (2014) Studies 1 and 2 Therm, Wave 1")
```



2. Joint Distribution of Feeling Thermometer and Policy Item

The conditional distribution of the feeling thermometer at every level of the SSM measure is also similar in both studies, despite the claim that the studies were drawn from distinct samples. The plot below shows the conditional distribution of the feeling thermometer at every level of the SSM item in each study and lists the marginal distribution of that SSM item.

```
plot.level <- function(study, level){
  data <- subset(lacour.reshaped, STUDY == paste0("Study ",study) & SSM_Level.1 == level)
  hist(data$Therm_Level.1, breaks=101, ylab=NA,
       main = paste0("Study ",study), cex=.5, axes=FALSE, xlim=c(0,100),
       xlab = paste0(
         "Therm | SSM = ", level,"; Pr(SSM=",level,")=",
               round(
                nrow(data)/nrow(subset(lacour.reshaped, STUDY == paste0("Study ",study)))
                  ,2)
                   )
       )
}
# Break up figures into multiples.
par(mfrow=c(2,2))
for(level in 1:2){
  for(study in 1:2){
    plot.level(study,level)
  }
}
```



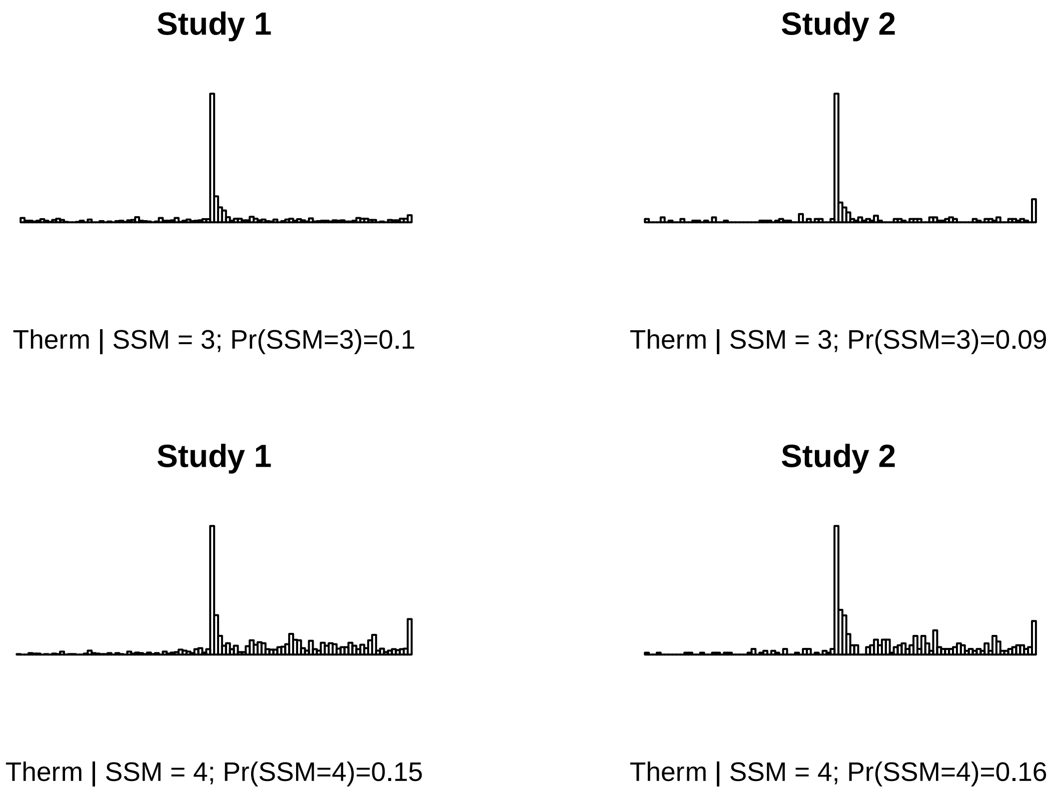
Study 1

Study 2

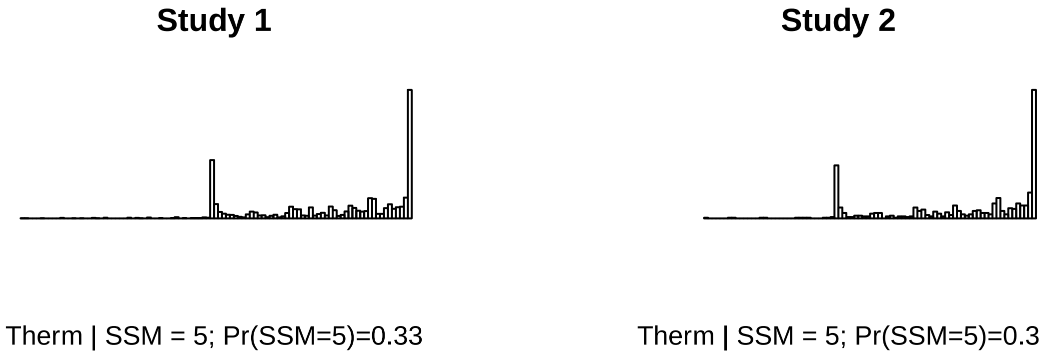
Study 1

Study 2

```r
par(mfrow=c(2,2))
for(level in 3:4){
  for(study in 1:2){
    plot.level(study,level)
  }
}
```



```
par(mfrow=c(2,2))
for(study in 1:2){
  plot.level(study,5)
}
plot.new()
plot.new()
```



3. High Reliability of the Feeling Thermometer

Feeling thermometers are notoriously unreliable survey items. That is, in a technical sense, subject's responses to feeling thermometers typically contain a fairly large amount of random measurement error. Measurement error should lead to attenuated correlations between subjects' wave 1 responses and their responses to the follow-up waves. However, these correlations are extremely strong in this dataset.

For this analysis and most others, we restrict our attention to the control group. The paper reports that the effects were heterogeneous by canvasser attributes, but canvasser indicators are not present in the replication data; this makes it difficult to account for the pattern of simulated treatment effects.

The thermometer readings at wave 1 and wave 2 for Study 1 are nearly perfectly correlated.

```
lacour.study1.controlgroup <- subset(lacour.reshaped,
           STUDY == "Study 1" & Treatment_Assignment == "No Contact")
cor(lacour.study1.controlgroup$Therm_Level.1,
    lacour.study1.controlgroup$Therm_Level.2, use = 'complete.obs')
```

```
## [1] 0.9975817
```

An early version of the paper notes that, in Study 1, the thermometers on subjects' screens were set to their wave 1 values in the second wave, providing a potential explanation for this pattern. Therefore, we restrict our attention to Study 2 for the remainder. In Study 2, the test-retest correlations remain extremely high.

```
lacour.study2.controlgroup <- subset(lacour.reshaped,
        STUDY == "Study 2" & Treatment_Assignment == "No Contact")
lacour.study2.therms <- lacour.study2.controlgroup[,c('Therm_Level.1',
        'Therm_Level.2', 'Therm_Level.3', 'Therm_Level.4')]
cor(lacour.study2.therms, use = 'complete.obs')
```

```
##               Therm_Level.1 Therm_Level.2 Therm_Level.3 Therm_Level.4
## Therm_Level.1     1.0000000     0.9720847     0.9588907     0.9692368
## Therm_Level.2     0.9720847     1.0000000     0.9313616     0.9413132
## Therm_Level.3     0.9588907     0.9313616     1.0000000     0.9343502
## Therm_Level.4     0.9692368     0.9413132     0.9343502     1.0000000
```

```
cor(lacour.study2.therms, use = 'pairwise.complete.obs')
```

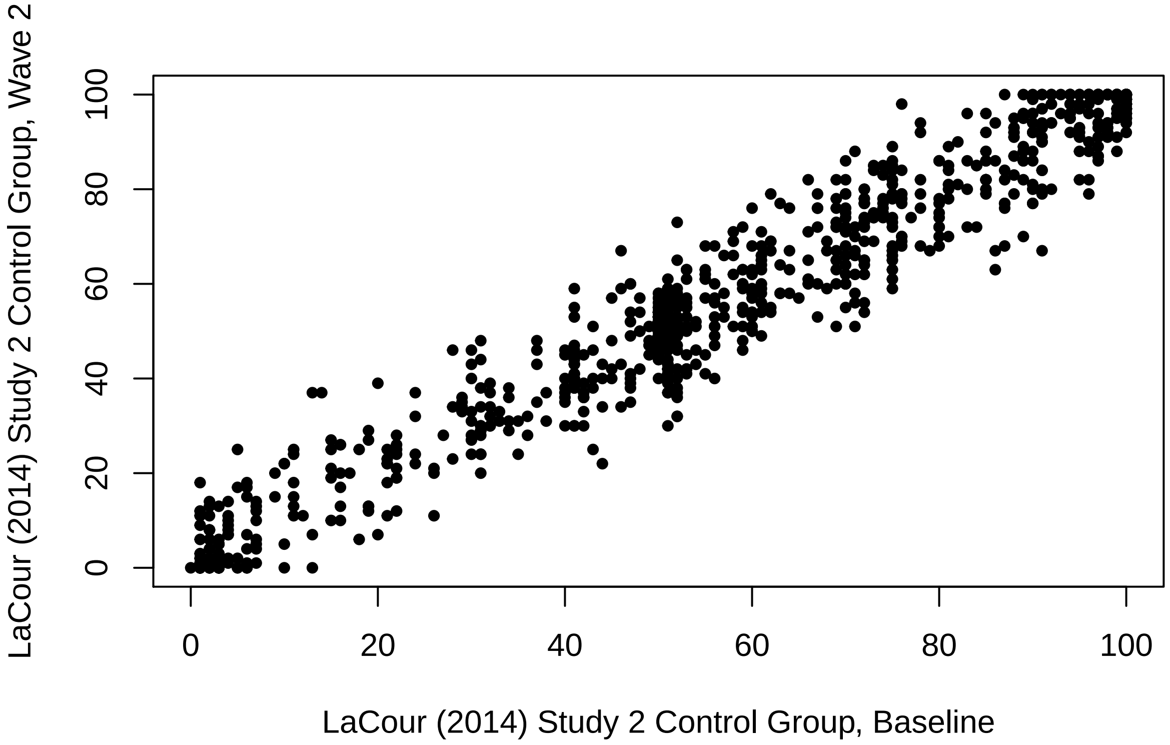
```
##               Therm_Level.1 Therm_Level.2 Therm_Level.3 Therm_Level.4
## Therm_Level.1     1.0000000     0.9734449     0.9594085     0.9709017
## Therm_Level.2     0.9734449     1.0000000     0.9308287     0.9436621
## Therm_Level.3     0.9594085     0.9308287     1.0000000     0.9343249
## Therm_Level.4     0.9709017     0.9436621     0.9343249     1.0000000
```

4. Distributions of Changes in Feeling Thermometers Highly Regular

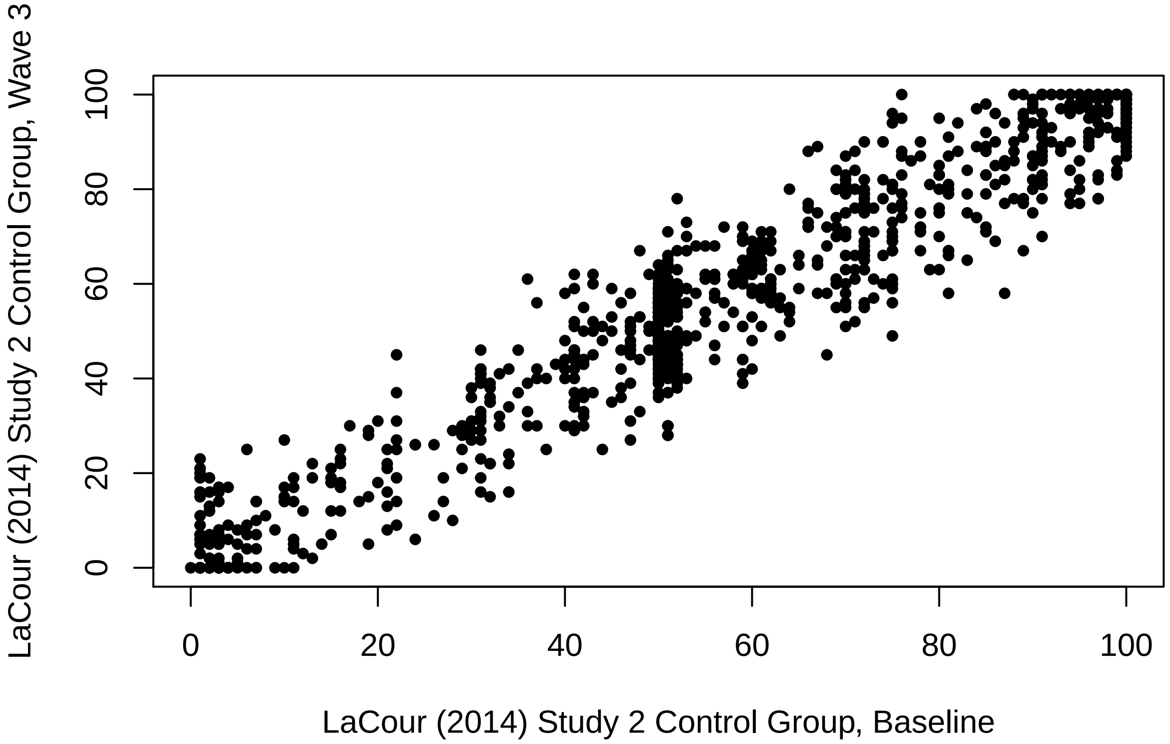
Our next few points illustrate several patterns consistent with the hypothesis that waves 2, 3, and 4 of the feeling thermometer in Study 2 were generated by adding normal noise to the baseline CCAP data and then truncating this noise at 0 on the lower and at 100 on the upper end.

First, not only are the feeling thermometers highly stable over time on average, but not one of the 3160 responses to this item markedly deviate from the baseline wave plus a normal distribution. In most survey response data we would expect to see at least one such deviation.

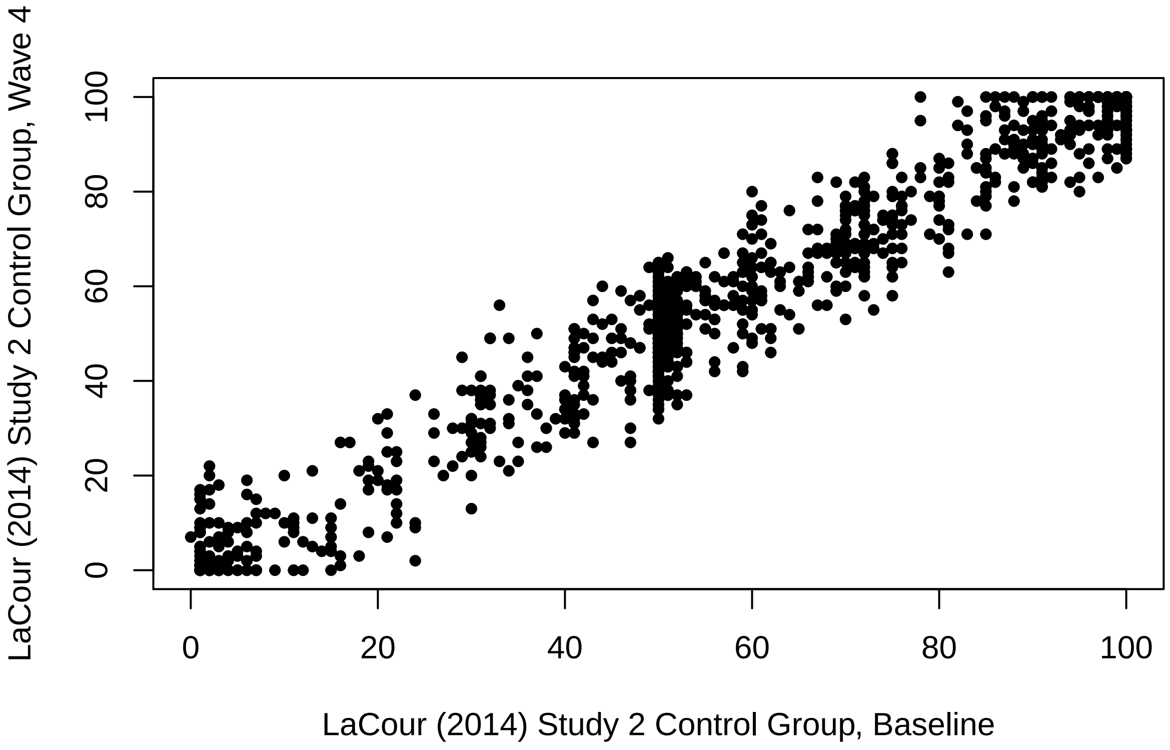
```
par(mfrow=c(1,1))
plot(lacour.study2.controlgroup$Therm_Level.1, lacour.study2.controlgroup$Therm_Level.2,
     pch=20, xlab = "LaCour (2014) Study 2 Control Group, Baseline",
     ylab = "LaCour (2014) Study 2 Control Group, Wave 2")
```



```
plot(lacour.study2.controlgroup$Therm_Level.1,
     lacour.study2.controlgroup$Therm_Level.3, pch=20,
     xlab = "LaCour (2014) Study 2 Control Group, Baseline",
     ylab = "LaCour (2014) Study 2 Control Group, Wave 3")
```



```
plot(lacour.study2.controlgroup$Therm_Level.1,
     lacour.study2.controlgroup$Therm_Level.4, pch=20,
     xlab = "LaCour (2014) Study 2 Control Group, Baseline",
     ylab = "LaCour (2014) Study 2 Control Group, Wave 4")
```



5. Follow-up Waves exhibit different heaping than Wave 1

In general, feeling thermometer responses show heaping patterns, with respondents being more likely to provide certain values. We see these patterns in the baseline data / the 2012 CCAP data – for example, respondents were especially likely to answer exactly at 50. However, the follow-up waves do not exhibit these patterns but instead appear to be offset by normally distributed shocks.

Heaping at 50 In Wave 1 but Not Follow-Up Waves

```
table(lacour.study2.controlgroup$Therm_Level.1 == 50)
```

```
##
## FALSE  TRUE
##   972   231
```

This pattern is expected in feeling thermometers. However, this pattern disappears in the subsequent waves.

```
table(lacour.study2.controlgroup$Therm_Level.2 == 50)
```

```
##
## FALSE  TRUE
##  1005    34
```

```
table(lacour.study2.controlgroup$Therm_Level.3 == 50)
```

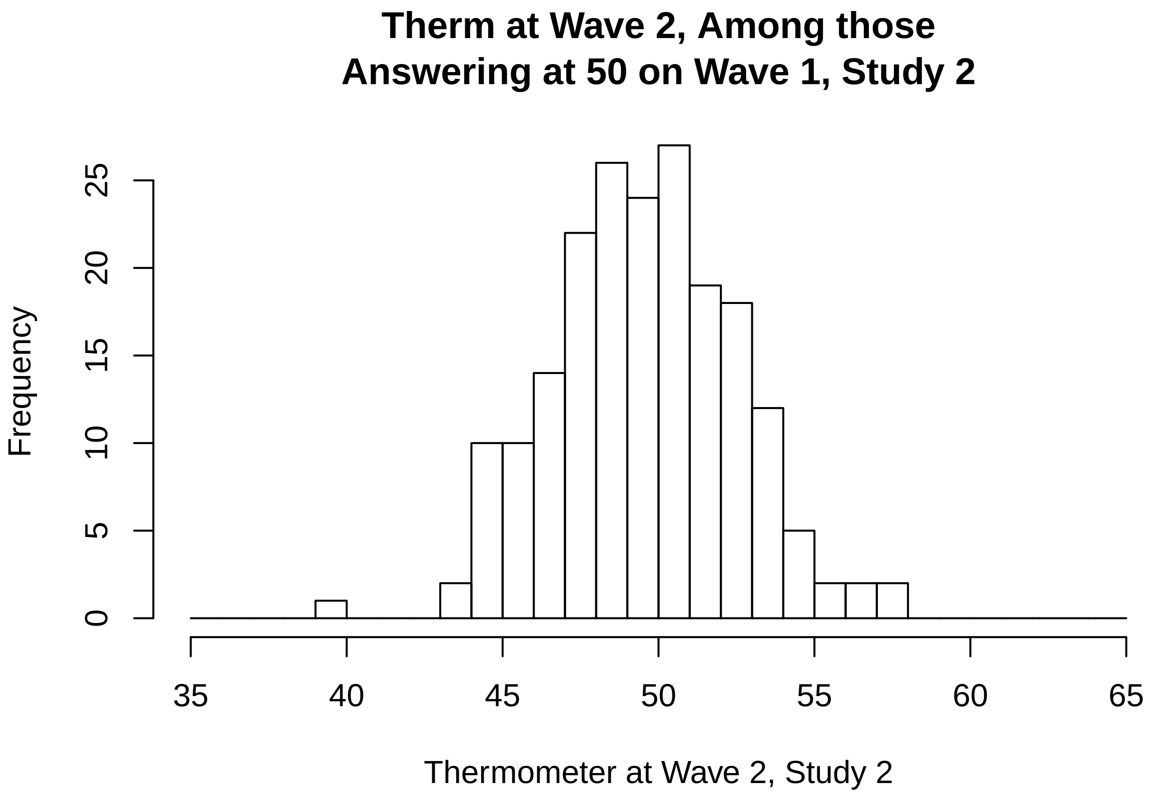
```
##
## FALSE  TRUE
##  1032    23
```

```
table(lacour.study2.controlgroup$Therm_Level.4 == 50)
```

```
##
## FALSE  TRUE
##  1046    20
```

Instead, everyone who answered at 50 previously is offest by a normally distributed shock, with respondents not showing any special preference for 50.

```
hist(lacour.study2.controlgroup$Therm_Level.2[lacour.study2.controlgroup$Therm_Level.1==50],
     breaks=seq(from=35,to=65,by=1),
     main = "Therm at Wave 2, Among those\nAnswering at 50 on Wave 1, Study 2",
     xlab = "Thermometer at Wave 2, Study 2")
```



No Heaping at 0 in Baseline Wave, Much in Follow-Up Wave

By contrast, only one respondent in the baseline wave answered exactly at 0, a believable result based on the CCAP data.

```
table(lacour.study2.controlgroup$Therm_Level.1 == 0)
```

```
##
## FALSE  TRUE
##  1202     1
```

However, many respondents answered exactly at 0 in follow-up waves.

```
table(lacour.study2.controlgroup$Therm_Level.2 == 0)
```

```
##
## FALSE  TRUE
##  1001    38
```

```
table(lacour.study2.controlgroup$Therm_Level.3 == 0)
```

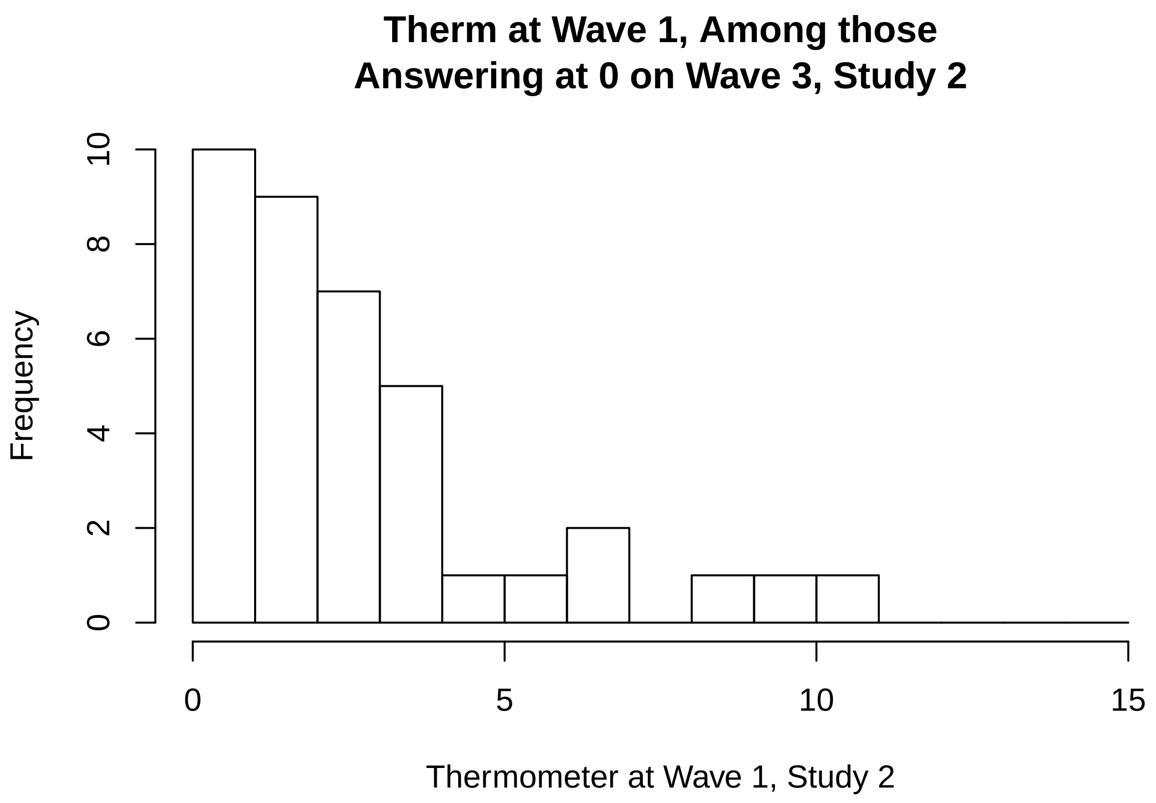
```
##
## FALSE  TRUE
##  1017    38
```

```
table(lacour.study2.controlgroup$Therm_Level.4 == 0)
```

```
##
## FALSE  TRUE
##  1021    45
```

This is consistent with the 0 responses being generated by truncation after those with low responses received shocks that put them outside the range of possible responses. For example, below are the wave 1 responses of the 38 subjects who answered at 0 in wave 3.

```
hist(lacour.study2.controlgroup$Therm_Level.1[lacour.study2.controlgroup$Therm_Level.3==0],
     breaks=seq(from=0,to=15,by=1),
     main = "Therm at Wave 1, Among those\nAnswering at 0 on Wave 3, Study 2",
     xlab = "Thermometer at Wave 1, Study 2")
```



6. Changes in the item format are unlikely to be responsible

One possibility that could explain the finding in Section 5 above is that the item format changed beginning with wave 2, yielding different patterns of heaping due to differences in how respondents could register their attitudes. Perhaps the clearest evidence that item format changes did not occur is that in a regression predicting the third wave, the second wave provides no information beyond what is present in the first wave. Even a small amount of non-random measurement error present in waves 2 and 3 due to a different item format than in wave 1 should lead to some patterns in these waves not present in wave 1.

```
summary(lm(Therm_Level.3 ~ Therm_Level.1 + Therm_Level.2, data=lacour.study2.controlgroup))
```

```
##
## Call:
## lm(formula = Therm_Level.3 ~ Therm_Level.1 + Therm_Level.2, data = lacour.study2.controlgroup)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.8569  -4.4977   0.5022   4.6343  26.0678
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.42017    0.61026   3.966 7.89e-05 ***
## Therm_Level.1  0.98551    0.04054  24.310  < 2e-16 ***
## Therm_Level.2 -0.02991    0.04074  -0.734    0.463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 8.017 on 907 degrees of freedom
##   (293 observations deleted due to missingness)
## Multiple R-squared:  0.9191, Adjusted R-squared:  0.919
## F-statistic:  5154 on 2 and 907 DF,  p-value: < 2.2e-16
```

Another way to appreciate this pattern is that the many respondents who answered at 50 at baseline but no longer heap at 50 in waves 2, 3, and 4 have uncorrelated wave 2, wave 3, and wave 4 responses. If the data were as reliable as the test-retest correlations suggested, we would expect subjects who were no longer enticed to answer exactly at 50 would show at least some consistent preference for one side of the 50 mark.

```
answered.at.50.baseline <- subset(lacour.study2.controlgroup, Therm_Level.1==50)
cor(cbind(answered.at.50.baseline$Therm_Level.2, answered.at.50.baseline$Therm_Level.3,
          answered.at.50.baseline$Therm_Level.4), use='complete.obs')
```

```
##             [,1]        [,2]       [,3]
## [1,]  1.00000000 -0.08401570 0.01385222
## [2,] -0.08401570  1.00000000 0.00451131
## [3,]  0.01385222  0.00451131 1.00000000
```

Were the item format to have changed, we may also expect the wave 2, 3, and 4 responses to correlate together at least as strongly with each other as with wave 1, with its different item format (unless the item format in wave 1 was considerably more reliable). However, each of the follow-up waves actually correlates more strongly with the baseline wave than it does with the other follow-up waves.

```
cor(lacour.study2.therms, use = 'complete.obs')
```

```
##               Therm_Level.1 Therm_Level.2 Therm_Level.3 Therm_Level.4
## Therm_Level.1     1.0000000     0.9720847     0.9588907     0.9692368
## Therm_Level.2     0.9720847     1.0000000     0.9313616     0.9413132
## Therm_Level.3     0.9588907     0.9313616     1.0000000     0.9343502
## Therm_Level.4     0.9692368     0.9413132     0.9343502     1.0000000
```

```
cor(lacour.study2.therms, use = 'pairwise.complete.obs')
```

```
##               Therm_Level.1 Therm_Level.2 Therm_Level.3 Therm_Level.4
## Therm_Level.1     1.0000000     0.9734449     0.9594085     0.9709017
## Therm_Level.2     0.9734449     1.0000000     0.9308287     0.9436621
## Therm_Level.3     0.9594085     0.9308287     1.0000000     0.9343249
## Therm_Level.4     0.9709017     0.9436621     0.9343249     1.0000000
```

Changes over time are also uncorrelated across waves.

```
lacour.study2.therm.changes <- lacour.study2.controlgroup[,c('Therm_Change.2',
      'Therm_Change.3', 'Therm_Change.4')]
cor(lacour.study2.therm.changes, use = 'complete.obs')
```

```
##                Therm_Change.2 Therm_Change.3 Therm_Change.4
## Therm_Change.2    1.000000000    0.004660893   -0.008939679
## Therm_Change.3    0.004660893    1.000000000    0.077151273
## Therm_Change.4   -0.008939679    0.077151273    1.000000000
```

```
cor(lacour.study2.therm.changes, use = 'pairwise.complete.obs')
```

```
##                Therm_Change.2 Therm_Change.3 Therm_Change.4
## Therm_Change.2    1.000000000   -0.003088789   -0.009142816
## Therm_Change.3   -0.003088789    1.000000000    0.065957305
## Therm_Change.4   -0.009142816    0.065957305    1.000000000
```

7. Endogenous takeup of treatment appears completely random

This experiment considered door-to-door canvassing. In door-to-door canvassing experiments, assignment to treatment is random and expected to be unrelated to baseline covariates. However, whether a voter can be successfully reached is endogenous, and typically (though not always) related to the outcome of interest (for discussion, see Table 1 in http://www.campaignfreedom.org/doclib/20110131_GerberandGreen2005.pdf). However, in LaCour (2014), we do not see signs of this pattern in either study. Below, we see no evidence of baseline differences between the groups receiving no contact, direct contact, or 'secondary' contact.

```
table(lacour.reshaped$Contact)
```

```
## 
##    Direct      None Secondary 
##       676     10085      1187
```

```
summary(lm(Therm_Level.1 ~ Contact, data=subset(lacour.reshaped, STUDY == "Study 1")))
```

```
## 
## Call:
## lm(formula = Therm_Level.1 ~ Contact, data = subset(lacour.reshaped, 
##     STUDY == "Study 1"))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -59.904 -10.387  -6.387  25.613  42.598 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        59.904      1.251  47.881   <2e-16 ***
## ContactNone        -1.517      1.291  -1.175    0.240    
## ContactSecondary   -2.502      1.570  -1.593    0.111    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 28.5 on 9504 degrees of freedom
## Multiple R-squared:  0.0002679,  Adjusted R-squared:  5.751e-05 
## F-statistic: 1.273 on 2 and 9504 DF,  p-value: 0.2799
```

```
summary(lm(Therm_Level.1 ~ Contact, data=subset(lacour.reshaped, STUDY == "Study 2")))
```

```
## 
## Call:
```

```
## lm(formula = Therm_Level.1 ~ Contact, data = subset(lacour.reshaped,
##     STUDY == "Study 2"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.887  -9.887  -6.887  26.113  42.915
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        58.1401     2.2796  25.504   <2e-16 ***
## ContactNone         0.7464     2.3674   0.315    0.753
## ContactSecondary   -1.0556     2.8407  -0.372    0.710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.56 on 2438 degrees of freedom
## Multiple R-squared:  0.0004259,  Adjusted R-squared:  -0.0003941
## F-statistic: 0.5194 on 2 and 2438 DF,  p-value: 0.5949
```

Restricting our attention to just the treatment group, we see the same lack of baseline differences between those the campaign successfully contacted, contacted indirectly through an acquaintance, or failed to contact.

```
summary(lm(Therm_Level.1 ~ Contact, data=subset(lacour.reshaped,
                                                STUDY == "Study 1" & TA != "C")))
```

```
##
## Call:
## lm(formula = Therm_Level.1 ~ Contact, data = subset(lacour.reshaped,
##     STUDY == "Study 1" & TA != "C"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.904 -11.151  -6.151  25.096  42.598
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         59.904      1.252  47.850   <2e-16 ***
## ContactNone         -1.753      1.361  -1.288    0.198
## ContactSecondary    -2.502      1.571  -1.592    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.52 on 4266 degrees of freedom
## Multiple R-squared:  0.0006014,  Adjusted R-squared:  0.0001329
## F-statistic: 1.284 on 2 and 4266 DF,  p-value: 0.2771
```

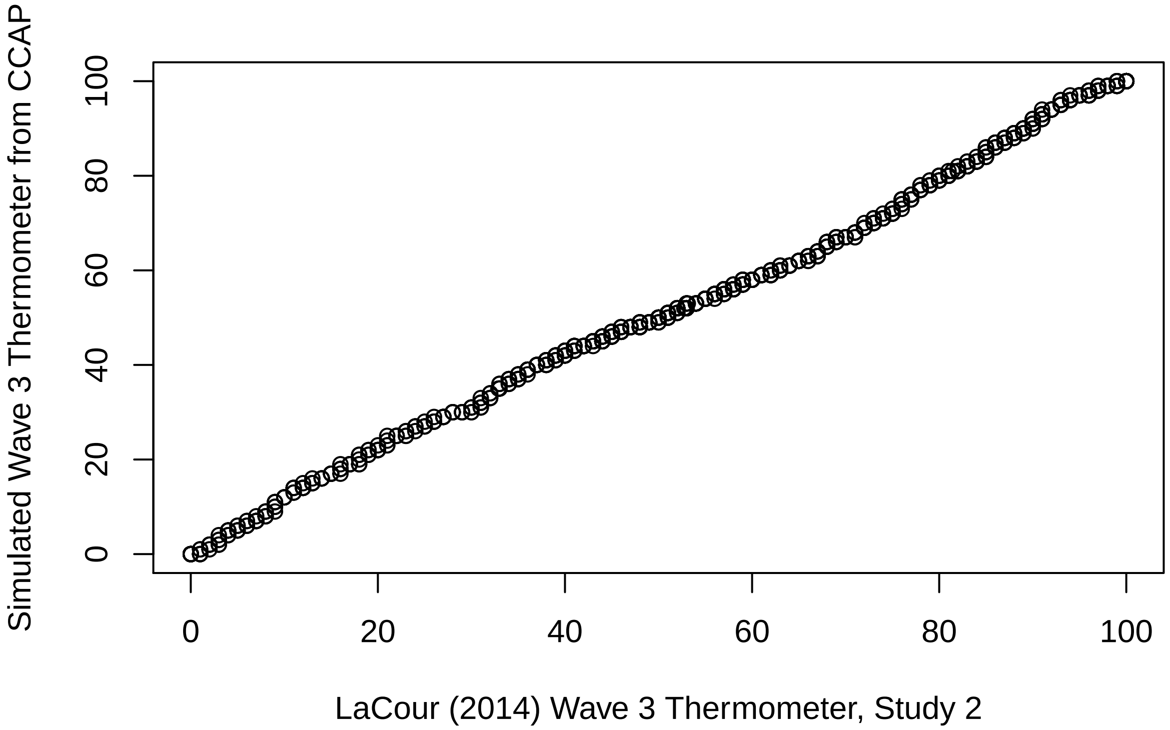
```
summary(lm(Therm_Level.1 ~ Contact, data=subset(lacour.reshaped,
                                                STUDY == "Study 2" & TA != "C")))
```

```
##
## Call:
## lm(formula = Therm_Level.1 ~ Contact, data = subset(lacour.reshaped,
##     STUDY == "Study 2" & TA != "C"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.396 -10.396  -6.268  25.604  42.915
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        58.140      2.267  25.649   <2e-16 ***
## ContactNone         2.256      2.480   0.910    0.363
## ContactSecondary   -1.056      2.825  -0.374    0.709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.4 on 1235 degrees of freedom
## Multiple R-squared:  0.002563,   Adjusted R-squared:  0.000948
## F-statistic: 1.587 on 2 and 1235 DF,  p-value: 0.205
```

8. Clear reproducibility based on simple data simulation

Following the simple procedure below generates patterns that are indistinguishable from the distribution in the data.

```
# Simulate the simple noise addition pattern we suspect.
therm3.simulated <- round(ccap.therm + rnorm(n = length(ccap.therm), mean = 0, sd = 8.4))
therm3.simulated[therm3.simulated<0] <- 0
therm3.simulated[therm3.simulated>100] <- 100

# Comparison of simulated and claimed real data - KS and QQ Plot
par(mfrow=c(1,1))
qqplot(therm3.simulated, lacour.study2.controlgroup$Therm_Level.3,
       xlab = "LaCour (2014) Wave 3 Thermometer, Study 2",
       ylab = "Simulated Wave 3 Thermometer from CCAP")
```

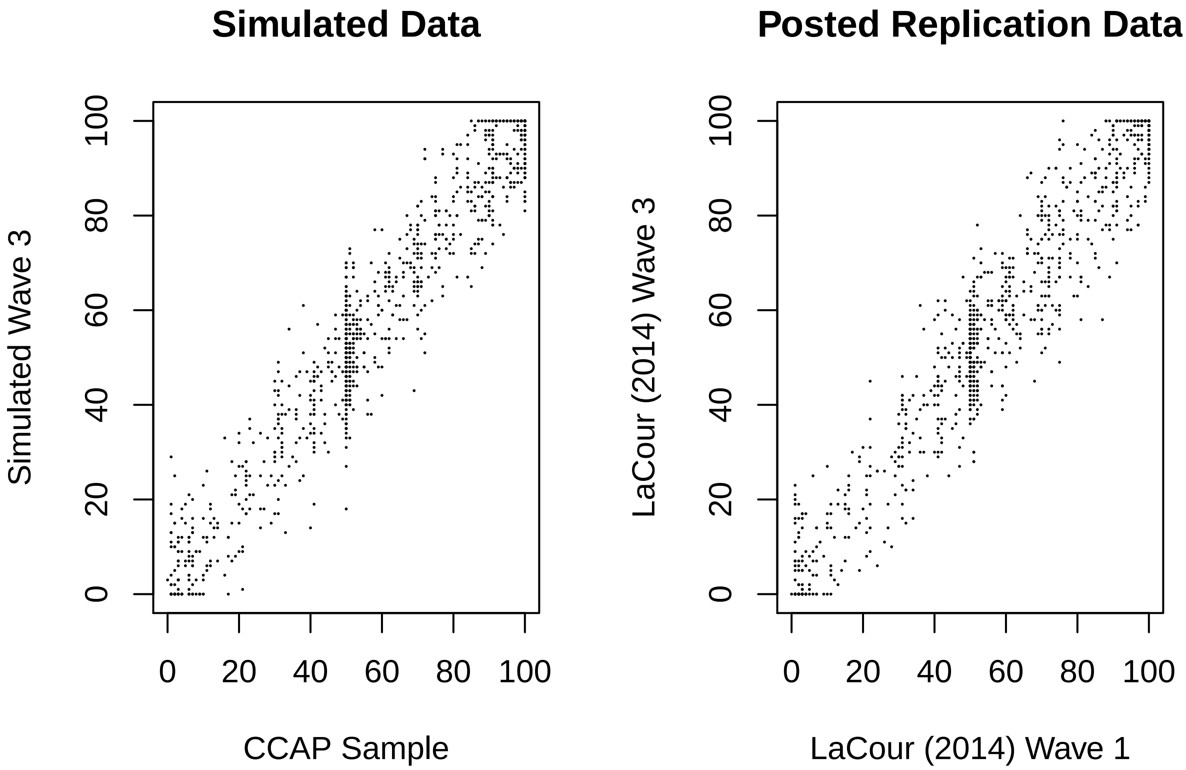


```
ks.test(therm3.simulated, lacour.study2.controlgroup$Therm_Level.3)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  therm3.simulated and lacour.study2.controlgroup$Therm_Level.3
## D = 0.0328, p-value = 0.2187
## alternative hypothesis: two-sided
```

Visually, a sample of the same size from the processed CCAP data bears a striking resemblance to these variables in LaCour (2014).

```
# Draw a sample from CCAP of the same size as Therm Level 3 to aid visual.
lacour.n <- sum(!is.na(lacour.study2.controlgroup$Therm_Level.3))
in.sample <- sample(lacour.n)
ccap.therm.sim.sample <- ccap.therm[in.sample]
therm3.simulated.sample <- therm3.simulated[in.sample]

par(mfrow=c(1,2))
plot(ccap.therm.sim.sample, therm3.simulated.sample, pch=16, cex = .2,
     main = "Simulated Data", xlab = "CCAP Sample", ylab = "Simulated Wave 3")
plot(lacour.study2.controlgroup$Therm_Level.1, lacour.study2.controlgroup$Therm_Level.3,
     pch=16, cex = .2, main = "Posted Replication Data",
     xlab = "LaCour (2014) Wave 1", ylab = "LaCour (2014) Wave 3")
```



Remaining Uncertainties

- We do not have access to the same-sex marriage question in CCAP, so we cannot evaluate the similarities of LaCour (2014)'s same-sex marriage question to the CCAP on that item.
- The claimed treatment effect was heterogeneous by canvasser attributes and the posted replication file does not have canvasser identifiers, so it is difficult to perform diagnostics on the responses of those assigned to treatment.
- The data for the abortion study reported at http://www.cis.ethz.ch/content/dam/ethz/special-interest/gess/cis/cis-dam/CIS_DAM_2015/Colloquium/Papers/LaCour_2015.pdf in LaCour (2015) is not currently publicly available.

References

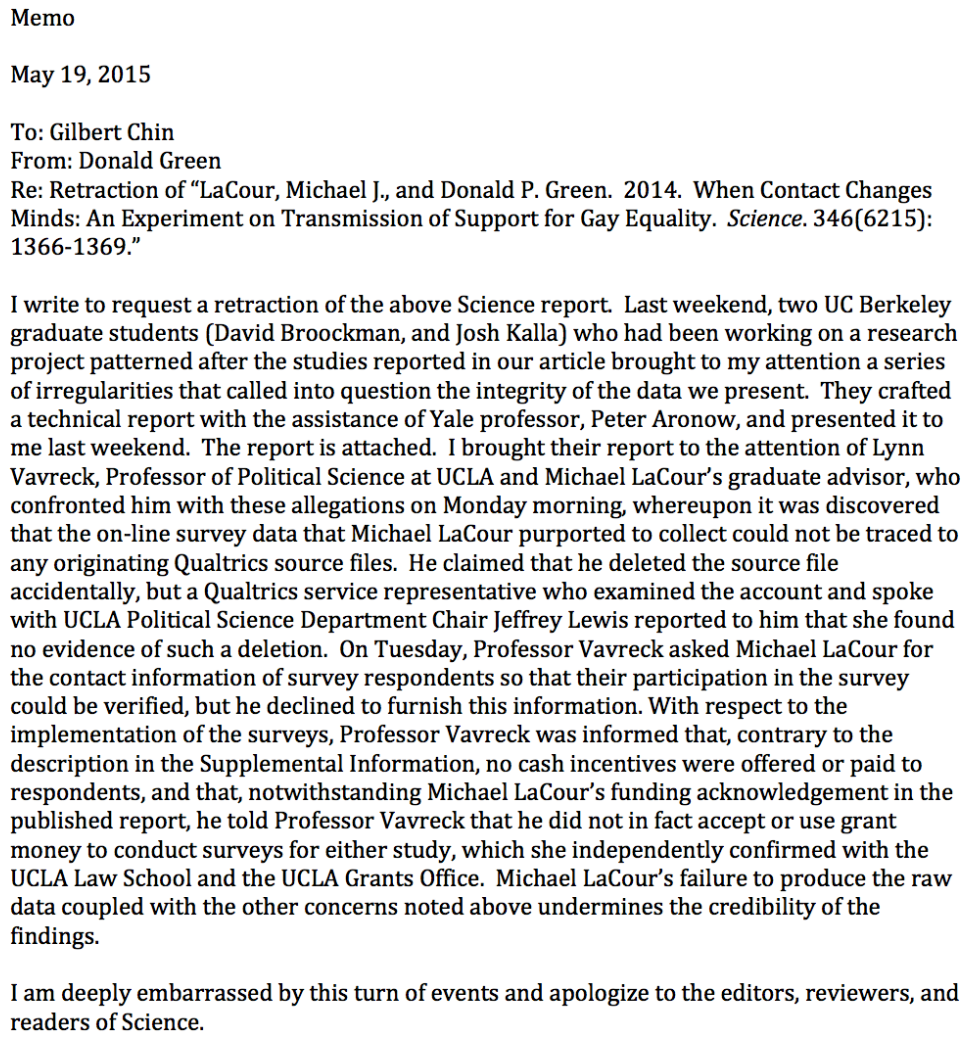
LaCour, Michael J. 2014-12-21. Political Persuasion and Attitude Change Study: The Los Angeles Longitudinal Field Experiments, 2013-2014: LaCour & Green (2014) Replication Files. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. http://doi.org/10.3886/E24287V1.

LaCour, Michael J. 2015. The Lasting Effects of Personal Persuasion. Working Paper, UCLA.

LaCour, Michael J. and Donald P. Green. 2014. When contact changes minds: An experiment on transmission of support for gay equality. Science 346(6215): 1366-1369.

Tesler, Michael. 2014. "Replication data for: Priming Predispositions and Changing Policy Positions", http://dx.doi.org/10.7910/DVN/26721, Harvard Dataverse, V2

Appendix - Green *Science* Retraction Request



Memo

May 19, 2015

To: Gilbert Chin
From: Donald Green
Re: Retraction of "LaCour, Michael J., and Donald P. Green. 2014. When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality. *Science*. 346(6215): 1366-1369."

I write to request a retraction of the above Science report. Last weekend, two UC Berkeley graduate students (David Broockman, and Josh Kalla) who had been working on a research project patterned after the studies reported in our article brought to my attention a series of irregularities that called into question the integrity of the data we present. They crafted a technical report with the assistance of Yale professor, Peter Aronow, and presented it to me last weekend. The report is attached. I brought their report to the attention of Lynn Vavreck, Professor of Political Science at UCLA and Michael LaCour's graduate advisor, who confronted him with these allegations on Monday morning, whereupon it was discovered that the on-line survey data that Michael LaCour purported to collect could not be traced to any originating Qualtrics source files. He claimed that he deleted the source file accidentally, but a Qualtrics service representative who examined the account and spoke with UCLA Political Science Department Chair Jeffrey Lewis reported to him that she found no evidence of such a deletion. On Tuesday, Professor Vavreck asked Michael LaCour for the contact information of survey respondents so that their participation in the survey could be verified, but he declined to furnish this information. With respect to the implementation of the surveys, Professor Vavreck was informed that, contrary to the description in the Supplemental Information, no cash incentives were offered or paid to respondents, and that, notwithstanding Michael LaCour's funding acknowledgement in the published report, he told Professor Vavreck that he did not in fact accept or use grant money to conduct surveys for either study, which she independently confirmed with the UCLA Law School and the UCLA Grants Office. Michael LaCour's failure to produce the raw data coupled with the other concerns noted above undermines the credibility of the findings.

I am deeply embarrassed by this turn of events and apologize to the editors, reviewers, and readers of Science.