**Abstract Title Page**
*Not included in page count.*

**Title:** Detecting anchoring-and-adjusting in survey scales

**Authors and Affiliations:** Joe McIntyre, Harvard Graduate School of Education

## Abstract Body

*Limit 4 pages single-spaced.*

### Background / Context:

*Description of prior research and its intellectual context.*

Proper survey design is essential to obtain reliable, replicable data from research subjects. One threat to inferences drawn from surveys is anchoring-and-adjusting. Tversky and Kahnemann (1974) observed that participants' responses to questions depended systematically on irrelevant information they received prior to answering. Epley and Gilovich (2006) provided further evidence that participants anchor on previous responses and then adjust insufficiently in answering the current question. Gehlbach and Barge (2012) demonstrated that anchoring and adjusting is more problematic in scales which are in uninterrupted blocks than in scales which are interrupted by items from other scales. They also showed that anchoring and adjusting can artificially inflate reliabilities as measured by Cronbach's alpha, while decreasing correlations with other scales.

### Purpose / Objective / Research Question / Focus of Study:

*Description of the focus of the research.*

In light of this research, it is important for survey designers to have tools that can detect anchoring-and-adjusting. Using non-experimental data, I will try to determine if respondents are anchoring and adjusting by looking for evidence that responses to adjacent items are more likely to be identical than they should be.

### Setting:

*Description of the research location.*

In this study I analyze scale data from two different settings. One set of scales comes from an online survey administered through the *Survey Monkey* website. The second set comes from a survey administered in person to undergraduates and graduates students at a private university in the Northeast.

### Population / Participants / Subjects:

*Description of the participants in the study: who, how many, key features, or characteristics.*

All participants were better educated than people in the country as a whole and one set of participants all had internet access. Thus, it is not appropriate to generalize my results to the population at large. However, this research is largely a proof-of-concept, designed to show that my question can be answered using non-experimental data. If it is successful, it will be important to apply this procedure to different sorts of scales and in different populations to see if there are differences in the amount of anchoring-and-adjusting.

### Intervention / Program / Practice:

*Description of the intervention, program, or practice, including details of administration and duration.*
(May not be applicable for Methods submissions)

### Significance / Novelty of study:

*Description of what is missing in previous work and the contribution the study makes.*

Although the phenomenon of anchoring-and-adjusting in questionnaire responses was established through a split-ballot randomized controlled trial, this rigorous approach will rarely be feasible for most survey researchers. Instead, it would be much more helpful to ascertain post-hoc whether anchoring-and-adjusting had occurred in a particular block of items. In developing a non-experimental approach to detecting anchoring and adjusting, I enable researchers to quickly and easily ascertain whether this bias has affected their data.

**Statistical, Measurement, or Econometric Model:**

*Description of the proposed new methods or novel applications of existing methods.*

If anchoring-and-adjusting is occurring, I expect that responses to adjacent items will be identical more frequently than they would if there were no anchoring-and-adjusting. I test this by attempting to determine frequently responses to adjacent items should be identical, and comparing this to how frequently I observe them to be. To determine how often responses should be identical, I employ a parametric bootstrap (Davison & Hinkley, 1999) in a categorical-indicator factor-analytic framework. I proceed as follows: first, I calculate $\hat{t}_{obs}$ in the sample as the proportion of responses to adjacent items which are identical. If respondents are anchoring-and-adjusting, I expect that $\hat{t}_{obs}$ will be large relative to values of $\hat{t}_{rep}$ calculated from datasets generated using the same scale but without anchoring-and-adjusting.

Next, I simulate new datasets to get a sense of what the distribution of the frequency of identical responses to adjacent items should be under a null-hypothesis of no anchoring-and-adjusting. Using the lavaan package (Rosseel, 2012) in R (R Core Team, 2013), I fit a saturated, categorical-indicators, factor-analytic model to a scale. I take this model to be a null distribution, $\hat{F}_0$. This distribution has the attractive properties that it is close to the observed data, in that the expected variance-covariance matrix and mean vector of data generated from this model will be identical to those in the observed data. At the same time, scales generated from $\hat{F}_0$ have no anchoring-and-adjusting in that responses to adjacent items are independent conditional on the respondent's scores on the latent variables. Put differently, if we were to rearrange the ordering of the items, we would expect to see identical data.

I proceed by randomly generating a large number of replicate datasets from $\hat{F}_0$. In each replicate dataset, I calculate $\hat{t}_{rep}$. I interpret these values of $\hat{t}_{rep}$ as the distribution of $\hat{t}$ under the null-hypothesis of no anchoring-and-adjusting. At this point I can derive a loose estimate of the relative "amount" of anchoring-and-adjusting by comparing the mean of $\hat{t}_{rep}$ to $\hat{t}_{obs}$. From Davison and Hinkley (1999), I can also derive a p-value associated with the null-hypothesis by taking

$$p_{H_0} = \frac{1+\#(\hat{t}_{rep} \geq \hat{t}_{obs})}{1+R},$$

where *R* is the number of replicate datasets.

Because this method is novel, I am concerned that it will not work well. Thus I test it on simulated data. I need to ensure that it performs acceptably, meaning that, given that a scale is

actually generated without anchoring-and-adjusting, it rejects approximately 5% of true null-hypotheses, and that $\hat{t}_{rep}$ is unbiased estimator for $\hat{t}_{obs}$. To do this, I randomly generate a large number of datasets from $\hat{F}_0$, as estimated above. I will apply the full procedure to each of these datasets generated without anchoring-and-adjusting, and ensure that it performs acceptably. If the procedure rejects more than 5% of the true null-hypotheses, I will use the distribution of the p-values, $p_{rep}$, to adjust my original p-value, by taking

$$p_{adj} = \frac{1+\#(p_{rep} \leq p_{obs})}{1+R_2},$$

where $R_2$ is the number of datasets I generated from $\hat{F}_0$.

I note in passing that, assuming that this method is appropriate, it is likely to be conservative, because insofar as anchoring-and-adjusting is occurring, the covariances of adjacent items are likely to be artificially inflated, meaning that datasets generated from $\hat{F}_0$ will tend to have higher values of $\hat{t}_{rep}$ than they would if they were generated from a distribution, $F$, with no anchoring-and-adjusting. Thus, any findings will be especially convincing.

Additionally, I will apply a permutation test to the scale as follows. First, I take $\hat{t}_{obs}$ to be the excess number of identical responses, or the difference between the observed number of responses to adjacent items which are identical and the model-implied number of responses to adjacent items which should be identical, as calculated using lavaan (Rosseel, 2012). Then, assuming that the excess number of identical responses is no greater for adjacent items than for non-adjacent items, I will randomly permute the order of the questions a large number of times, each time computing $\hat{t}_{rep}$. I control for the fact that question order is not randomly assigned by taking the excess number of identical responses as a test-statistic rather than the simple number of identical responses. As before, I compute

$$p_{H_0} = \frac{1+\#(\hat{t}_{rep} \geq \hat{t}_{obs})}{1+R}.$$

Note that this approach should be doubly conservative because not only will anchoring-and-adjusting increase the model-implied number of identical responses for adjacent items by increasing their covariances, it will also increase the observed number of identical responses to non-adjacent items because anchoring on a response anchor to, e.g., item 1 will slightly increase the probability of selecting that same response anchor on item 3. Thus, it will be harder to reject the null-hypothesis, even if it is false.

**Usefulness / Applicability of Method:**
*Demonstration of the usefulness of the proposed methods using hypothetical or real data.*

This approach should enable a researcher to take any scale data and find evidence of anchoring-and-adjusting, and to get a sense of its magnitude. This will make it easier to determine whether certain scale designs elicit more anchoring-and-adjusting than others, and whether some populations are more likely to engage in anchoring-and-adjusting than others. This is of substantive interest in and of itself, but it will also help researchers to identify possible threats to the validity of their inferences. I will make my code freely available to interested researchers,

and will consider wrapping it in an R package, if there is sufficient interest. The basic analyses take no more than two minutes to run for a seven-item scale with 1167 respondents, making it a reasonably easy analysis to run, especially compared to the effort required to design and run a good experiment. At the very least, this approach can help generate hypotheses about anchoring-and-adjusting, which can then be investigated experimentally.

**Research Design:**
*Description of the research design.*
(May not be applicable for Methods submissions)

**Data Collection and Analysis:**
*Description of the methods for collecting and analyzing data.*

Both the *Survey Monkey* and the university surveys employed a split-ballot design, in which half of the participants were randomly assigned to see scales with items phrased as statements with Likert style response agree-disagree (A-D) anchors (e.g. disagree, neither agree nor disagree, and agree), while others were assigned to see the same scales with items phrased as questions with either construct-specific or frequency-based response anchors.

**Findings / Results:**
*Description of the main findings with specific details.*

I find evidence of anchoring-and-adjusting on most scales (see table 1). As predicted, I find more evidence of anchoring-and-adjusting on A-D scales, and on uninterrupted scales. The one scale for which there is no evidence of anchoring-and-adjusting is an interrupted scale, and the other interrupted scale shows only marginal evidence of anchoring-and-adjusting. To demonstrate how my proposed methods work, I present more detailed analyses for one of the scales, an agreeably-acquiescing scale from the *Survey Monkey* survey. I find substantial, statistically significant evidence of anchoring and adjusting on the agreeably-acquiescing scale both using the parametric bootstrap (see figure 1) and the permutation test (see figure 2). I also find that the methods I employ are appropriate in that the estimator is unbiased when the null-hypothesis is true (see figure 3), and p-values for a true null-hypothesis are approximately uniformly distributed (see figure 4). The p-values reported in table 1 have been corrected for the slight non-uniformity using the technique from Davison and Hinkley (1999) described above.

**Conclusions:**
*Description of conclusions, recommendations, and limitations based on findings.*

It is possible to find evidence of anchoring-and-adjusting using non-experimental approaches. There is more evidence of anchoring-and-adjusting on uninterrupted scales than on interrupted scales. Further research is needed to determine what sorts of scales elicit more anchoring-and-adjusting, and which sorts of populations anchor-and-adjust more frequently. If it turns out that anchoring-and-adjusting is common in survey scales, it is also important to determine what effects it has on the validity of inferences drawn from scale data.

# Appendices

*Not included in page count.*

## Appendix A. References

*References are to be in APA version 6 format.*


Davison, A.C. & Hinkley, D.V. (1999) *Bootstrap methods and their application.* Cambridge, UK: Cambridge University Press.

Epley, N. & Gilovich, T. (2006). The anchoring and adjusting heuristic: Why adjustments are insufficient. *Psychological Science, 17*(4), 311-318.

Gehlbach, H. & Barge, S. (2012). Anchoring and adjusting in questionnaire responses. *Basic and Applied Social Psychology, 34*(5), 417-433.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 79*, 344-354.

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/.

## Appendix B. Tables and Figures

*Not included in page count.*

Table 1: Incidence of anchoring and adjusting on selected scales taken from 1167 participants.

| Scale | Items | Type | Observed[1] | Expected[2] | p-value[3] |
|---|---|---|---|---|---|
| Agreeability | 7 | C-S, uninterrupted | .36 | .33 | <.001 |
| Extraversion | 8 | C-S, uninterrupted | .25 | .22 | <.001 |
| Social Perspective Taking[4] | 7 | Frequency, uninterrupted | .41 | .38 | .06 |
| Social Perspective Taking[4] | 7 | A-D, uninterrupted | .46 | .34 | <.001 |
| Satisfaction | 5 | C-S, uninterrupted | .65 | .61 | .01 |
| Satisfaction | 5 | A-D, uninterrupted | .72 | .64 | <.001 |
| Involvement | 5 | C-S, uninterrupted | .31 | .21 | <.002 |
| Involvement | 5 | A-D, uninterrupted | .37 | .25 | <.001 |
| Fit and engagement, 1[5] | 7 | A-D and C-S, interrupted | .15 | .15 | .73 |
| Fit and engagement, 2[5] | 7 | A-D and C-S, interrupted | .13 | .14 | .79 |

[1] Observed proportion of identical responses to adjacent items

[2] Mean bootstrapped proportion of identical responses to adjacent items

[3] p-value associated with the null-hypothesis of no anchoring-and-adjusting

[4] Social perspective taking scales were administered on a survey of graduates and undergraduates at a private northeastern university

[5] Fit and engagement scales were spread across two pages; each entry represents one page

Figure 1: A histogram of $\hat{t}_{rep}$ from a parametric bootstrap with vertical lines at $\mu_{\hat{t}_{rep}}$ and $\hat{t}_{obs}$ for 1000 replications

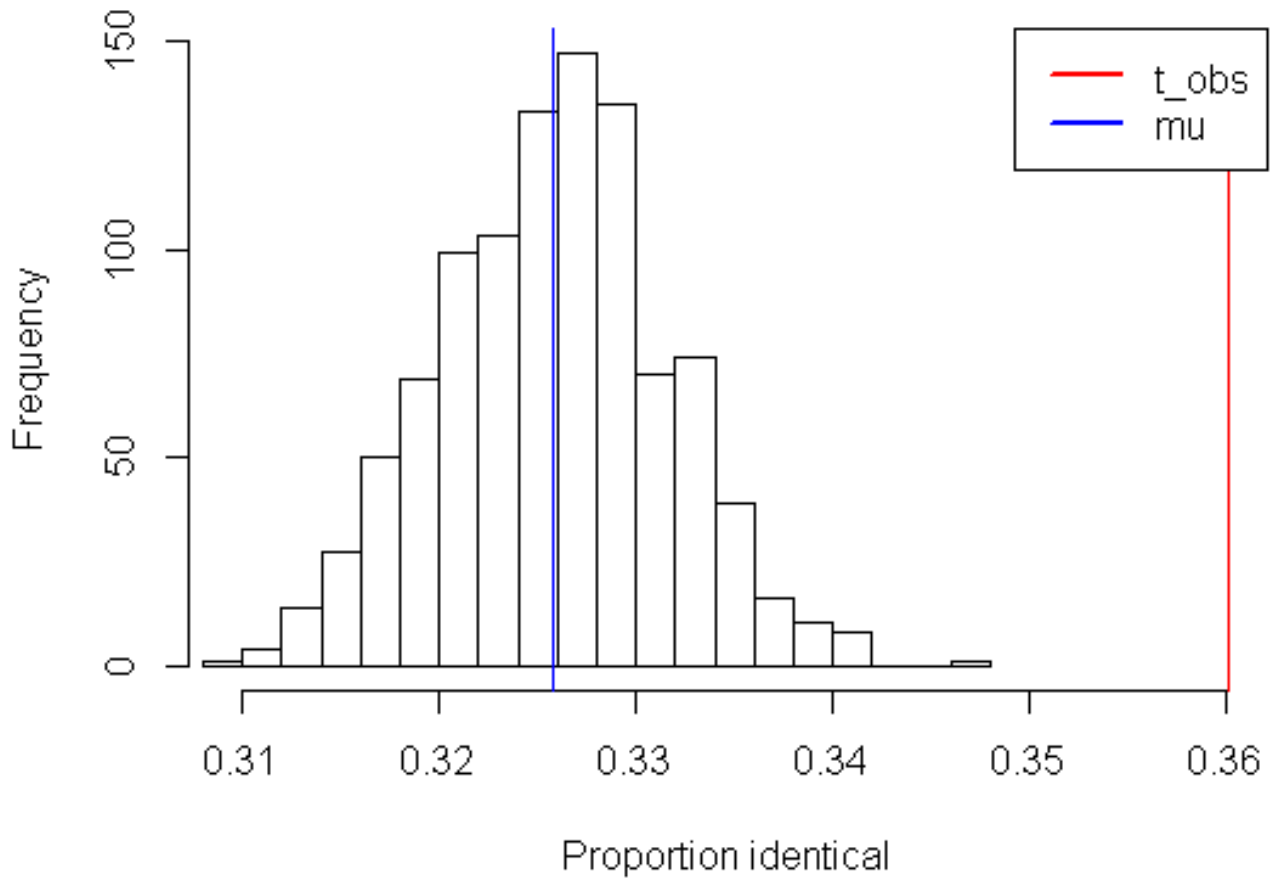

Figure 2: A histogram of $\hat{t}_{rep}$ from a permutation test with vertical lines at $\mu_{\hat{t}_{rep}}$ and $\hat{t}_{obs}$ for 1000 replications

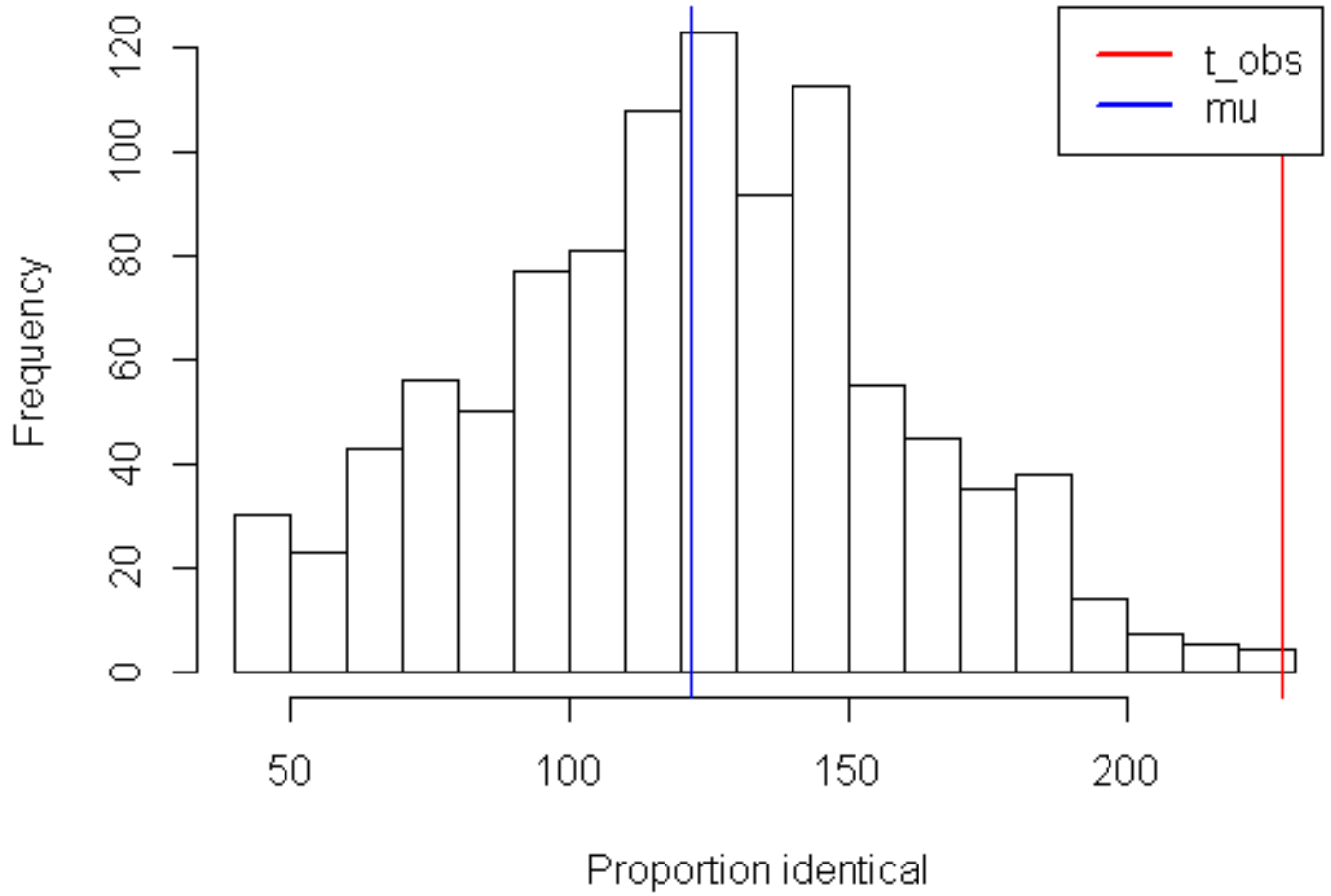

Figure 3: A histogram of $\hat{t}_{rep} - \mu_{t_{rep}}$ from a series of 1000 bootstraps with vertical lines at the observed difference and the mean of the bootstrapped differences.

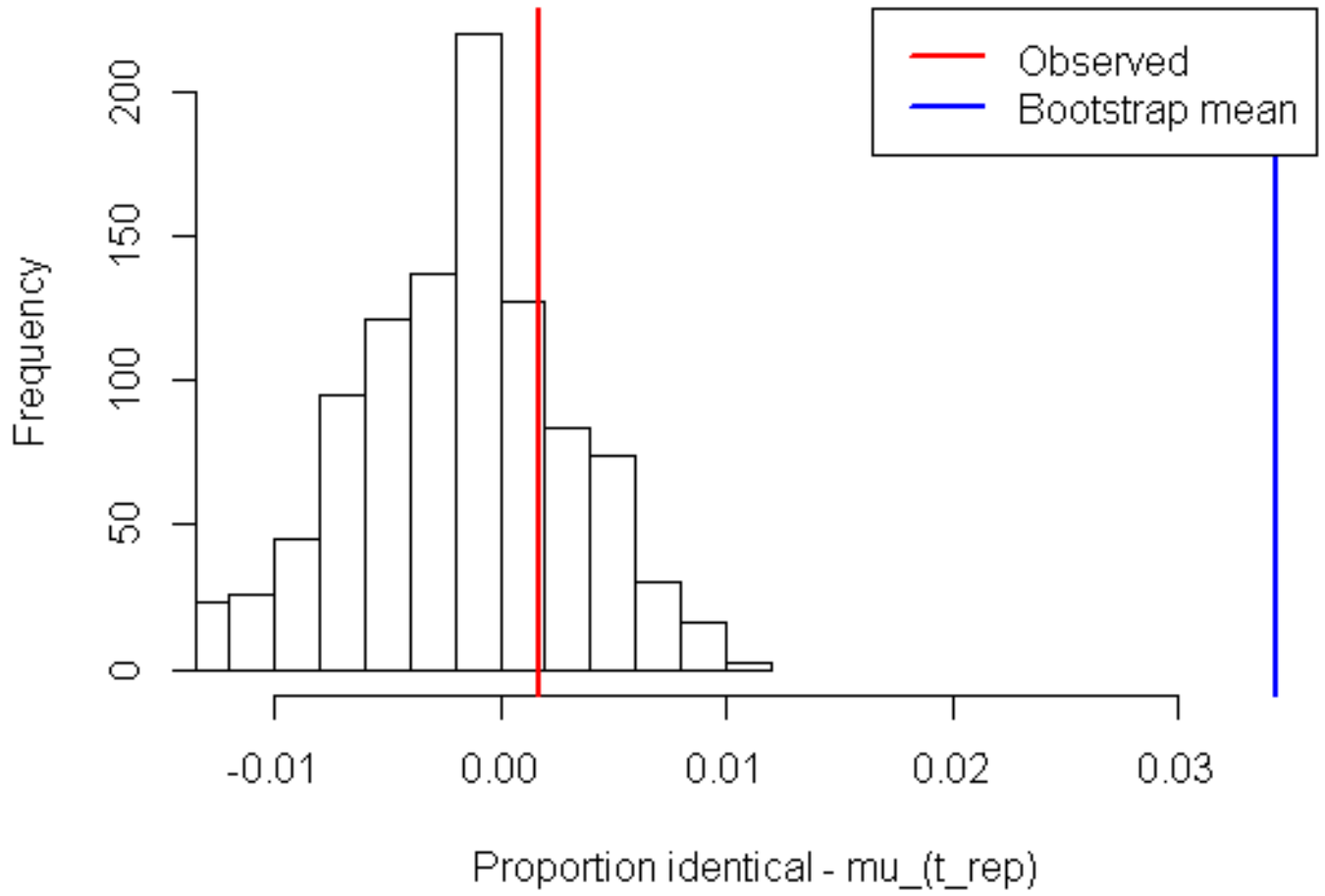

Figure 4: A histogram of p-values obtained from bootstrapped samples under a true null-hypothesis.