Time, Money, and Morality.
Psychological Science 0956797613506438.

C. J. Schwarz
Department of Statistics and Actuarial Science, Simon Fraser University
cschwarz@stat.sfu.ca
January 2, 2014

Contents

1 Introduction 2
2 Experiment 1 3
3 Experiment 2 6
4 Experiment 3 9
5 Experiment 4 9
6 Summary 11
1 INTRODUCTION

Abstract

Gino and Mogilner (2013) conducted a study to investigate the impact of priming subjects to think about “time” vs. thinking about “money” on the extent and amount of cheating in a series of 4 experiments.

While the overall omnibus analyses are correctly done, the authors did not adjust for multiple comparisons when performing the pairwise comparisons. In many cases, the adjusted comparisons fail to show evidence of a difference in the proportion who cheated or the amount of cheating, contrary to the authors’ unadjusted results.

1 Introduction

Gino and Mogilner (2013)\(^1\) did an interesting series of experiments on the impact of priming subjects to think about “time” vs. thinking about “money” on the extent and amount of cheating. This article attracted much media attention, including an article in the Economist\(^2\).

The authors ran 4 different experiment as described in their paper and summarized below. Briefly, subjects were primed in various ways to think about time or to think about money or think about a neutral topic. Then the subjects completed puzzles and self-reported their scores on the puzzles. The subjects believed that their self-reported scores were anonymous, but the authors were able to actually compare the self-reported scores with the actual scores to measure the proportion of subjects who cheated and the amount of cheating.

The experimental designs were either a single-factor (Experiments 1 and 4) or two-factor (Experiments 2 and 3) completely randomized designs. Standard chi-square tests or ANOVA were used for the omnibus tests of no difference in the proportion who cheated or the amount of cheating across the experimental factors. I was able to reproduce these omnibus results.

However, the authors then compared the proportions or means among the pairs of treatment levels without adjusting for multiple comparisons. For example, following an ANOVA, it is customary to perform a Tukey-adjustment when looking at pairwise differences in means; the authors did not make any adjustments.

The authors also failed to provide estimates of effect size and measures of precision for effect sizes making it difficult to identify the size of the effects.

---


detected.

R code and results of the reanalysis of the authors’ data are available at [http://www.stat.sfu.ca/~cschwarz/CourseNotes/Reanalysis](http://www.stat.sfu.ca/~cschwarz/CourseNotes/Reanalysis).

# Experiment 1

The methods are detailed in Gino and Mogilner (2013). Briefly, subjects were randomized to one of three conditions (money prime, time prime, or no prime (control)). The prime was a series of scrambled-sentences task where the subjects were exposed to time-related words, money-related words, or only neutral words.

Then all subjects were presented with a “numbers” game along with an envelope that contained $20 and two pieces of paper. One paper was a work sheet with no identifying information with 20 (apparently random) matrices each consisting of a set of 12 three-digit numbers (e.g. 4.78). Subjects need to find two numbers per matrix that added to 10 (e.g. 4.78 + 5.22). There was insufficient time to complete all the matrices.

After working through the worksheets, subjects recorded their score on the second sheet (with no identifying information), discarded the worksheets, and kept $1 for each correct answer.

Because there was no apparent identifying information anywhere on the sheets, subjects had both the incentive and opportunity to overreport this performances to earn or money. In fact, all sheets were coded through the (apparent) random numbers, so the authors could actually match the actual score with the reported score to measure the extent and amount of cheating.

The authors concluded:

“The percentage of participants who cheated varied across conditions, $\chi^2(2, N = 98) = 14.61, p = .001$ (see Fig. 1); participants were more likely to cheat in the money condition (87.5%, 28/32) than in either the control condition (66.7%, 22/33), $\chi^2(1, N = 65) = 3.97, p < .05$, or the time condition (42.4%, 14/33) $\chi^2(1, N = 65) = 14.44, p < .001$. Also, participants were less likely to cheat in the time condition than in the control condition, $\chi^2(1, N = 66) = 3.91, p < .05$. The extent of cheating also varied across conditions, $F(2, 95) = 5.09, p = .008, \eta^2_p = .10$. Simple contrasts revealed that participants cheated more in the money condition ($M = 4.41, SD = 4.25$) than in both the control condition ($M = 2.76, SD = 3.96; p =$...
and the time condition ($M = 1.55, SD = 2.41; p = .002$). The difference between the time and control conditions did not reach statistical significance ($p = .18$).”

They also included a graph of the results as shown in Figure 1:

![Figure 1: Figure 1 from Gino and Mogilner (2013) reporting the results of Experiment 1.](image)

When we redid the analysis to compare the proportion of participants that cheated among the three conditions, we did obtain the same $\chi^2$ test-statistic value and the same $p$-value as the authors. This $p$-value indicates that there was evidence of a difference in the proportion who cheated among the three experimental conditions, but does NOT indicate where the difference may lie. The authors then did followup pairwise-comparisons to examine if there were differences in the proportion of cheated among the pairs of conditions, but did NOT ADJUST the analysis of these comparisons to account for the multiple-testing problem.

These pairwise-comparisons must be adjusted for multiple-testing in an analogous way as is done in testing for differences in pairwise means in Analysis of Variance problems. A Tukey-like adjustment can be done by casting the problem as a generalized-linear model problem using logistic regression to compare the proportion who cheated in each group (refer to R code) giving:

| Prime lsmeans' | .group Prime logit.P.cheat SE df asymp.LCL asymp.UCL |
|---------------|------------------------|--------|----------------|-----------------|
| B Time       | -0.305 0.352 NA -0.996 0.385 |

©2013 Carl James Schwarz
2 EXPERIMENT 1

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>SE</th>
<th>df</th>
<th>z.ratio</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB Control</td>
<td>0.693</td>
<td>0.369</td>
<td>NA</td>
<td>-0.031</td>
<td>1.417</td>
</tr>
<tr>
<td>A Money</td>
<td>1.946</td>
<td>0.535</td>
<td>NA</td>
<td>0.898</td>
<td>2.994</td>
</tr>
</tbody>
</table>

$'Prime pairwise differences'$

<table>
<thead>
<tr>
<th></th>
<th>log(OR)</th>
<th>SE</th>
<th>df</th>
<th>z.ratio</th>
<th>p.value</th>
<th>OR.lcl</th>
<th>OR.ucl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Money - Control</td>
<td>1.253</td>
<td>0.65</td>
<td>NA</td>
<td>1.928</td>
<td>0.131</td>
<td>3.500</td>
<td>0.980</td>
</tr>
<tr>
<td>Money - Time</td>
<td>2.251</td>
<td>0.64</td>
<td>NA</td>
<td>3.517</td>
<td>0.001</td>
<td>9.500</td>
<td>2.709</td>
</tr>
<tr>
<td>Control - Time</td>
<td>0.999</td>
<td>0.51</td>
<td>NA</td>
<td>1.957</td>
<td>0.123</td>
<td>2.714</td>
<td>0.998</td>
</tr>
</tbody>
</table>

p values are adjusted using the tukey method for 3 means

The compact-letter display (seen at the left of the first section of the output above) indicates that after adjustment for multiple-comparisons, there is NO evidence of a difference in the proportion of cheating (at $\alpha = 0.05$) between each of the primes and the control condition.

Why the contradiction? The authors did not adjust for multiple-comparisons, and, rather than reporting the actual $p$-value of their comparison (which is standard practise), simply reported $p < 0.05$, hiding the fact that any comparison was actually quite marginal and that an adjustment for multiple-comparisons would likely not show evidence of a difference.

Their Figure 1 is also deceptive because it does not present a measure of precision. If 95% confidence intervals are drawn on each of the bars in their Figure 1, the apparent conclusions are quite different as shown in Figure 2 where the confidence intervals show that evidence for differences in the proportion who cheated are not clear cut.

The authors then compared the amount of cheating among the three conditions. Again, the key problem is failing to control for multiple-comparisons when doing the pairwise comparisons following the single-factor completely randomized design ANOVA. The revised results can be computed in $R$ giving

$'Prime lsmeans'$

<table>
<thead>
<tr>
<th>group</th>
<th>Prime lsmean</th>
<th>SE</th>
<th>df</th>
<th>lower.CL</th>
<th>upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Money</td>
<td>4.41</td>
<td>0.64</td>
<td>3.138</td>
<td>5.682</td>
</tr>
<tr>
<td>AB</td>
<td>Control</td>
<td>2.76</td>
<td>0.63</td>
<td>1.508</td>
<td>4.012</td>
</tr>
<tr>
<td>B</td>
<td>Time</td>
<td>1.55</td>
<td>0.63</td>
<td>0.298</td>
<td>2.802</td>
</tr>
</tbody>
</table>

$'Prime pairwise differences'$

<table>
<thead>
<tr>
<th></th>
<th>estimate</th>
<th>SE</th>
<th>df</th>
<th>t.ratio</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Money - Control</td>
<td>1.65</td>
<td>0.899</td>
<td>95</td>
<td>1.835</td>
<td>0.164</td>
</tr>
<tr>
<td>Money - Time</td>
<td>2.86</td>
<td>0.899</td>
<td>95</td>
<td>3.181</td>
<td>0.006</td>
</tr>
<tr>
<td>Control - Time</td>
<td>1.21</td>
<td>0.892</td>
<td>95</td>
<td>1.356</td>
<td>0.368</td>
</tr>
</tbody>
</table>
3 Experiment 2

In the second experiment, subjects were randomized to one of 4 treatments in a two-factor completely randomized design. One factor was the type of Prime (money or time) where subjects in a treatment with the particular prime searched for lyrics involving time or money. The second factor was how the
test involving the same types of matrices in Experiment 1 were to be evaluated
(Assessment), with subjects being told that it was either an Intelligence test or
a Personality test. The self-reported scores for each subject were again matched
against the actual sheets from each subject. The authors stated:

“A 2 (prime) × 2 (assessment) analysis of variance was conducted on
extent of cheating, calculated as the difference between participants’
reported and actual performance on the numbers game. The results
revealed a marginal main effect of prime; participants in the money
condition cheated more than those in the time condition, $F(1, 138) = 2.77, p = .099$. As predicted, this effect was qualified by a significant
interaction, $F(1, 138) = 3.99, p < .05, \eta^2_p = .03$ (see Table 1 for
information about the percentage of people who cheated and the
extent of cheating in each condition). Only when the game was
framed as an intelligence test did thinking about money lead to
greater cheating than thinking about time, $F(1, 138) = 6.69, p = .01$. When the game was framed as a personality test, there was no
difference in cheating between the money and time conditions, $F < 1$. In fact, participants primed with money cheated less when they
thought the game assessed their personality than when they thought
it assessed their intelligence, $F(1, 138) = 4.58, p = .03$. There was
no such difference among those primed with time, $F < 1$.”
A summary of the raw data was presented in the authors’ Table 1 which enabled me to reconstruct their analyses.

The model for this experiment is a two-factor completely-randomized ANOVA or, in a standard shorthand notation:

\[ Y = \text{Prime} + \text{Assessment} + \text{Prime:Assessment} \]

where \(Y\) is the amount of cheating, \(\text{Prime}\) is the main effect of the Prime factor; \(\text{Assessment}\) is the main effect of the Assessment factor, and the \(\text{Prime:Assessment}\) term represents the interaction effect of the the factors on the mean response. I was able to reproduce the test-statistics and \(p\)-values for the test of no main effect of Prime and the test for no interaction effect between the two factors. However, for the latter test, the authors again simply reported that \(p < 0.05\) rather than reporting the actual \(p\)-value which was just under 0.05.

The authors then do a series of contrasts comparing pairs of treatments. Again, they failed to adjust the results for multiple-comparisons. The adjusted results (from \(R\)) are:

\begin{verbatim}
$'Prime:Assessment lsmeans'
group Prime Assessment lsmean SE df lower.CL upper.CL
  B Time Intel 0.27 0.213 138 -0.152 0.692
  AB Money Pers 0.42 0.204 138 0.016 0.824
  AB Time Pers 0.49 0.207 138 0.080 0.900
  A Money Intel 1.03 0.199 138 0.637 1.423

$'Prime:Assessment pairwise differences'
estimate SE df t.ratio p.value
Money, Intel - Time, Intel 0.76 0.292 138 2.606 0.049
Money, Intel - Money, Pers 0.61 0.285 138 2.140 0.146
Money, Intel - Time, Pers 0.54 0.287 138 1.881 0.241
Time, Intel - Money, Pers -0.15 0.295 138 -0.508 0.957
Time, Intel - Time, Pers -0.22 0.297 138 -0.740 0.881
Money, Pers - Time, Pers -0.07 0.291 138 -0.241 0.995

p values are adjusted using the tukey method for 4 means
\end{verbatim}

Now all but one of the pairwise comparisons so now evidence of a difference in the mean response, save for the comparison of the means of the \((\text{Money, Intelligence})\) treatment vs. the \((\text{Time, Intelligence})\) treatment, where the \(p\)-value is again just barely < 0.05. This is contrast to the two results reported by the authors which imply a song evidence of a difference.
In summary, while the technical analyses performed by the authors are correct, they again failed to adjust for multiple-comparisons when examining the pairwise comparisons and reported results in a way that was misleading (e.g. reporting $p < .05$ when the $p$-value is just below 0.05).

4 Experiment 3

The design of Experiment 3 was similar to that of Experiment 2, i.e. a two-factor completely randomized design. The factors were Prime (Money or Time), and Mirror (Present, Absent) where the mirror was used to increase self-awareness of the subjects.

The analysis proceeds similarly. The authors once again did not adjust for the multiple-comparisons after the preliminary analysis of variance result. Fortunately, the results were strong enough that the conclusions did not change even after doing an adjustment for multiple-comparisons.

5 Experiment 4

Experiment 4 was a single-factor completely randomized design where subjects were randomized to one of three Primes (Money, Control, or Time). The proportion who cheated and the amount of cheating on a word-scramble quiz was recorded.

In the analysis of the proportion who cheater, the authors report:

“Cheating. We observed the same pattern of results for cheating, $\chi^2(2, N = 213) = 16.44, p < .001$: Participants were more likely to cheat in the money condition (73.3%, 55/75) than in either the control condition (57.4%, 39/68), $\chi^2(1, N = 143) = 4.04, p = .044$, or the time condition (40.0%, 28/70), $\chi^2(1, N = 145) = 16.44, p < .001$. Participants were less likely to cheat in the time condition than in the control condition, $\chi^2(1, N = 138) = 4.16, p = .041$.”

As in Experiment 1, the authors failed to adjust for the multiple comparisons following the overall chi-square test. The adjusted results (from R) are:

\[
\begin{array}{cccccc}
\text{Prime} & \text{lsmeans} \\
\text{group} & \text{Prime} & \text{logit.P.cheat} & \text{SE} & \text{df} & \text{asymp.LCL} & \text{asymp.UCL} \\
\end{array}
\]
5 EXPERIMENT 4

$\log(OR)$ | SE | df | z.ratio | p.value | OR.lcl | OR.ucl
---|---|---|---|---|---|---|
Money - Control | 0.715 | 0.358 | NA | 1.997 | 0.113 | 2.045 | 1.013 | 4.126
Money - Time | 1.417 | 0.357 | NA | 3.965 | 0.000 | 4.125 | 2.048 | 8.310
Control - Time | 0.702 | 0.346 | NA | 2.029 | 0.105 | 2.017 | 1.024 | 3.974

‘Prime pairwise differences’

p values are adjusted using the tukey method for 3 means

and a plot of the results is shown in Figure 4. As in Experiment 1, there is no evidence of a difference in the proportion who cheated in the pairwise comparisons relative to the Control treatment.

![Results of Experiment 4 - Proportion cheating](image)

Figure 4: Results of experiment 4 from Gino and Mogilner (2013) with 95% confidence intervals added.

The participants also reported a self-reflection score. The authors reported the results of the analysis of this score as:

**Self-reflection.** Participants’ reported self-reflection varied by condition, $F(2, 210) = 12.42, p < .001, \eta^2_p = .11$ (see Table 3). In particular, reported levels of self reflection were lower in the money condition compared with both the control condition ($p = .001$) and the
time condition \( (p < .001) \). Participants reported greater self reflection in the time condition than in the control condition \( (p = .024) \).

As in previous experiment, the authors failed to adjust their pairwise comparisons for the multiple comparisons. The adjusted results (from \( R \)) are:

<table>
<thead>
<tr>
<th>.group</th>
<th>Prime lsmean</th>
<th>SE</th>
<th>df</th>
<th>lower.CL</th>
<th>upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Money</td>
<td>3.75</td>
<td>0.123</td>
<td>210</td>
<td>3.508</td>
</tr>
<tr>
<td>B</td>
<td>Control</td>
<td>4.22</td>
<td>0.129</td>
<td>210</td>
<td>3.966</td>
</tr>
<tr>
<td>B</td>
<td>Time</td>
<td>4.63</td>
<td>0.127</td>
<td>210</td>
<td>4.380</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>.group</th>
<th>Prime pairwise differences</th>
<th>estimate</th>
<th>SE</th>
<th>df</th>
<th>t.ratio</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Money - Control</td>
<td>-0.47</td>
<td>0.178</td>
<td>210</td>
<td>-2.640</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>Money - Time</td>
<td>-0.88</td>
<td>0.177</td>
<td>210</td>
<td>-4.981</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Control - Time</td>
<td>-0.41</td>
<td>0.181</td>
<td>210</td>
<td>-2.265</td>
<td>0.063</td>
<td></td>
</tr>
</tbody>
</table>

p values are adjusted using the tukey method for 3 means

After adjusting for multiple comparisons, there is no evidence of a difference in the mean self-reflection score between the Control and Time treatments.

The authors also performed an Analysis of Covariance to investigate if the differences in the amount cheating related to the Prime factor were reduced after adjusting for the self-reflection score. Unfortunately, without the raw data, I am unable to verify the results. In the interests of Reproducible Research, it is important that the raw data be made available over and above the summary statistics so that readers may reproduce results.

6 Summary

The key problem in this paper is that the authors failed to adjust the results from pairwise comparisons to account for the multiple testing following the omnibus test. By not making an adjustment, the authors will have increased their Type I (false positive) error rate. Indeed, as seen above, many of the pairwise comparisons show no evidence of a difference in the proportion who cheat or the amount of cheating contrary to the reported results of the author.

The authors also reported some results as simply \( p < 0.05 \) which was misleading (in the least) given that the actual \( p \)-value was just slightly below 0.05.
It is standard practice to report the actual $p$-value rather than simply $p < 0.05$ to avoid misleading the readers.

Similarly, their Figures and Tables in the their paper do not report confidence intervals for the marginal estimates, nor confidence intervals for effect sizes. The authors reporting of results is actually contrary to the journal policy.

“In previous editions of the journal’s Submission Guidelines, authors were advised: “Effect sizes should accompany major results. When relevant, bar and line graphs should include distributional information, usually confidence intervals or standard errors of the mean.”