Numerical Inconsistencies in Dubois, Rucker & Galinsky (2012) v2

By Aaron Charlton | email: abcharlton@protonmail.com

Note about version 2: David Dubois provided participant data from Study 3 which explained some of the data anomalies. Specifically, it showed that he consistently downrounded throughout Study 3 (e.g. 2.357 -> 2.35). This provided an innocuous explanation for GRIM inconsistencies in Study 3 and prompted me to reanalyze the other studies with downrounding in mind.

Overview:
When summary and test statistics reported in a study are based on the same set of participant data, they should have certain numerical consistencies among them. For example, the f-value for an ANOVA can be recalculated from the sample size, mean and standard deviation of each group. The same holds true for other types of summary and test statistics. In my forensic analysis of Dubois, Rucker & Galinsky (2012), I discovered substantial numerical inconsistency among summary and test statistics. The authors informed me that David Dubois did data collection and analysis for this article.

Contents

P-value errors

P-values don't match test statistics. This was pointed out to me by the authors. This problem was also pointed out in the replication. It is a fairly universal problem throughout the paper. The authors were informed of the issue by Leif Nelson when the paper first came out. The authors attempted to issue a correction but the journal declined. At the time, the authors generated a list of the errors along with their corrections but they have not yet shared this document with me. David Dubois indicated that he could get the document to me in Fall of 2022 or later when he is able to travel internationally.

Detailed list of the errors: David Dubois went through and identified p-value errors in 2022, put it in a spreadsheet and shared it with me. A link to the spreadsheet is found in the supplement section below.

Effect size errors

Effect sizes should be a function of the means and standard deviations. That is not the case here. This problem was pointed out in the replication. The replicators recalculated every effect size based on the reported summary and test statistics and arrived at different numbers from the original authors. See the supplementary materials of the replication: https://osf.io/f35w9/.

Reported in Dubois et al. (2012)	Recalculated in the replication (Ziano et al. 2022)	Notes
$\eta^2p = .05$	$\eta^2p = .022$	
$\eta^2p = .10$	$\eta^2p = .055$	
d = 1.10	d = 1.51	
d = 0.46	d = .90	
d = 0.65	d = .63	This difference is slight and could be due to rounding error

I have not checked other effect sizes in Dubois et al. (2012) for this problem.

Impossible means (GRIM issues)

Only certain means reported to two decimals are possible when the sample size for each cell is less than 100 and each value that goes into the mean is an integer value. In the following studies, exact cell sizes are unknown but the total sample size is reported. Because we don't know the exact cell sizes, I've had to run the GRIM Test for the entire range of plausible cell sizes. We also don't know whether downrounding was used (as we know occurred in Study 3 because we have the participant data), so I've had to test the means for regular rounding, downrounding, and a mix of the two types of rounding.

Note that each mean must (1) work for a plausible sample size and (2) all means within a given cell must work for the same cell size. Uneven missingness within a cell can cause some discrepancy #2 but would not affect #1.

R code used for GRIM testing is in the appendix. The means and SDs are included in a supplement at the end of this report.

For more about this method, see the GRIM Test paper.

Study 1

While exact cell sizes aren't reported, we know that the total sample size is 183 in a 9-cell design (average cell size = 20.33). Further, we know that all five dimension variables are reported as single-item 7-point (1-7) scales.

Table 1. Only 82% of means reported in Study 1 (Table 1) are possible for cell sizes between 15 and 25

```
|rounding_type          | How many means work individually?|
|:--------------------|--------------------------------:|
|works_downrounding     |                              0.73|
|works_mixed_rounding   |                              0.91|
|works_regular_rounding |                              0.82|
```
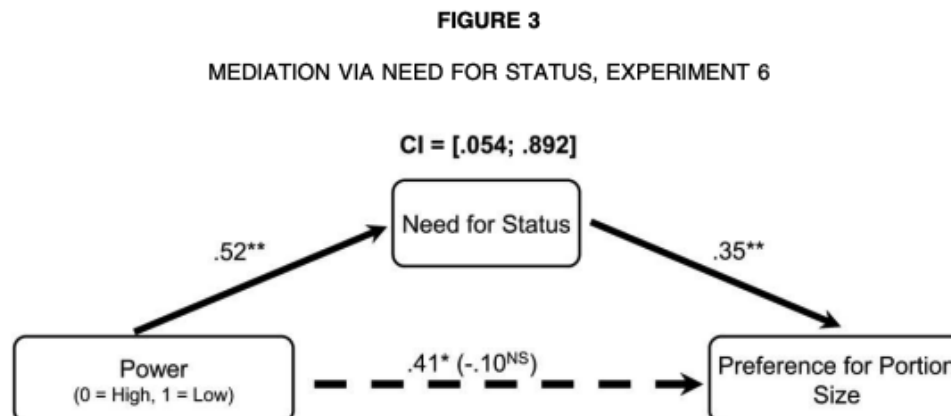
Note: Because we know that D. Dubois consistently used downrounding for Study 3, I checked all the means for regular rounding, downrounding, and a mixture of the two types of rounding.

Table 2. When cell sample size is constrained to be equal across all 5 DVs, 40% to 80% of reported means are possible for cell sizes between 15 and 25

```
|size   |product  | works_regular_rounding| works_downrounding| works_mixed_rounding|
|:------|:--------|---------------------:|----------------:|-------------------:|
|large  |coffee   |                   0.4|              0.6|                 0.6|
|large  |pizza    |                   0.4|              0.6|                 0.6|
|large  |smoothie |                   0.6|              0.4|                 0.6|
|medium |coffee   |                   0.4|              0.6|                 0.6|
|medium |pizza    |                   0.4|              0.4|                 0.6|
|medium |smoothie |                   0.8|              0.4|                 0.8|
|small  |coffee   |                   0.6|              0.6|                 0.6|
|small  |pizza    |                   0.6|              0.8|                 1.0|
|small  |smoothie |                   0.4|              0.4|                 0.4|
```

Note: Because we know that D. Dubois consistently used downrounding for Study 3, I checked all the means for regular rounding, downrounding, and a mixture of the two types of rounding.

Impossible mediation model in Study 6

FIGURE 3

MEDIATION VIA NEED FOR STATUS, EXPERIMENT 6

CI = [.054; .892]

Need for Status

.52**   .35**

Power
(0 = High, 1 = Low)   .41* (-.10^NS)   Preference for Portion Size

NOTE.—* = $p < .05$; ** = $p < .01$.

Severe asymmetry. The first problem is severe asymmetry of the reported confidence interval. The results section and the mediation model figure report a CI of [.054; .892] but the ab path point estimate is .52 * .35 = .182 based on the numbers given in the model. That means that the point estimate is 5.5x as far from the upper limit as it is from the lower limit. Slight asymmetry is normal but this amount of asymmetry is not possible.

Impossible full mediation. The second problem is that the indirect effect (a*b = .182) is far too weak to completely wipe out and change direction of the c = .41 total effect of X on Y. The formula for calculating the residual X->Y effect (c') is c' = c - a x b. Thus based on a, b, and c given in the model, c' should be .41 - .52*.35 = .228. The c' given in the chart (-.10 NS) is mathematically impossible. This is true regardless of whether the coefficients are standardized because c and a*b are both transformed in exactly the same way when standardized (SD_X/SD_Y), meaning the transformations cancel each other out when directly comparing a*b with c. You can see David Kenny's power estimator tool to test this: MedPower and you can also just look at other mediation model diagrams to see that this relationship holds up consistently.

Test statistic errors in Study 2

The standard deviations for each mean in Study 2 is implausibly large for a -1/0/1 three-level variable. If we accept them and recreate the ANOVA anyway, we see that the test statistic is much weaker than in the paper and nonsignificant: $F(2,139) = 2.845$, $p = .06$. All of the comparisons are also non-significant. Original test in the paper: $F(1, 135) = 10.54$, $p < .01$).

Results for Study 2 from https://statpages.info/anova1sm.html

Enter your summary data here...

Group Name	N (count)	Mean	Std. Dev. ⌄
Group 1	47	.32	.72
Group 2	47	.03	.79
Group 3	48	-.04	.82
Group 4			
Group 5			
Group 6			
Group 7			
Group 8			
Group 9			
Group 10			

Desired confidence level for post-hoc confidence intervals: 95

Compute

ANOVA Table...

Source of Variation	Sum of Squares	d.f.	Variance	F	p
Between Groups:	3.4451	2	1.7226	2.8451	0.0615
Within Groups:	84.1578	139	0.6055		
Total:	87.6029	141			

Post-hoc tests...

```
Tukey HSD Post-hoc Test...
Group 1 vs Group 2: Diff=-0.2900, 95%CI=-0.6702 to 0.0902, p=0.1711
Group 1 vs Group 3: Diff=-0.3600, 95%CI=-0.7383 to 0.0183, p=0.0658
Group 2 vs Group 3: Diff=-0.0700, 95%CI=-0.4483 to 0.3083, p=0.8996
```

May 30, 2022

Supplement

Appendix: R code for GRIM testing

```r
library(readr)
library(tidyverse)

########### Study 1
supersized <- read_csv("super-size-me-statistics - means_sds.csv")
means <- supersized %>%
  pivot_longer(cols = c(`High status`,Respected,Honest,Nice,Attractive),
          names_to = "dv", values_to ="value") %>%
  filter(statistic == "Mean") %>%
  select(value)

# check if all plausible sample size x mean combinations work individually
all_combos <- tibble(
  id = rep(1:45,11),
  mean = rep(means$value,11),
  size = rep(c("small","medium","large"),15*11),
  product = rep(rep(c("coffee","pizza","smoothie"), each = 3),5*11),
  sample_size = rep(15:25, each = 45),
  works_regular_rounding = trunc(sample_size * (mean + .005)) - trunc(sample_size * (mean -.005)),
  works_downrounding = trunc(sample_size * (mean + .01)) - trunc(sample_size * (mean)),
  works_mixed_rounding = trunc(sample_size * (mean + .01)) - trunc(sample_size * (mean -.005))
)

# how many means work individually for any plausible sample size?
s1_works_individually <- all_combos %>% group_by(id,mean,size,product) %>%
  summarise(works_regular_rounding = ifelse(sum(works_regular_rounding) > 1,1,0),
        works_downrounding = ifelse(sum(works_downrounding) > 1,1,0),
        works_mixed_rounding = ifelse(sum(works_mixed_rounding) > 1,1,0)) %>%
  pivot_longer(cols = starts_with("works"),
          names_to = "rounding_type",
          names_prefix = "wk",
          values_to = "works") %>%
  group_by(rounding_type) %>%
  summarise(`How many means work individually?` = mean(works))

kable(s1_works_individually, digits = 2)

# How many of the 5 DV means in each cell work for the same sample size?
s1_works_groups <- all_combos %>%
  group_by(mean, size,product,sample_size) %>%
  summarise(works_regular_rounding = mean(works_regular_rounding),
        works_downrounding = mean(works_downrounding),
```

```
        works_mixed_rounding = mean(works_mixed_rounding)) %>%
  group_by(size,product) %>%
  summarise(works_regular_rounding = max(works_regular_rounding),
        works_downrounding = max(works_downrounding),
        works_mixed_rounding = max(works_mixed_rounding))
library(knitr)
kable(s1_works_groups)
```