Letter to the Editors of Psychological Science: Resolving Inconsistencies With Data Gleaning: Regarding Bauer and Ariely (2021)

A recent Expression of Concern in this journal (Bauer and Ariely, 2021) was motivated by concerns that reported statistics had important errors/inconsistencies and that the original data was unavailable so the published results could not be verified. While looking over the paper in question (Heyman and Ariely, 2004), I noticed three things of interest.

First, the reported number of errors is incorrect. The expression of concern indicated that there were 13 errors, but repeating the *statcheck* analysis (Rife, Nuijten, & Epskamp, 2016) indicates 17 errors. A hand check of the manuscript indicates 19 errors (*statcheck* seemingly treats two of these cases as being due to rounding and thus not errors). I found that *statcheck* was unable to properly process a PDF copy of the article, and that copying the text of the article from the PDF resulted in missing symbols that may have confused the *statcheck* algorithm. I got the best results by copying text from the HTML version of the manuscript. There may be other peculiarities for different versions of *statcheck*; so, as suggested in the *statcheck* manual, it is important to manually verify *statcheck* calculations.

Second, the errors may be due to a simple description mistake. The reported tests seem to be based on contrasts from an ANOVA that used all groups to estimate pooled variance. With one degree of freedom in the numerator, a contrast test statistic can be treated as an *F*-value or as a *t*-value; and the tests are equivalent, with $F=t^2$. I suspect that the software program reported *t*-values for each test, but Heyman and Ariely (2004) reported them as *F*-values. For example, the reported test "$F(1, 84)=2.52$, $p=.014$" is inconsistent because the given *F*-value should have $p=.116$. If we interpret the test as actually being "$t(84)=2.52$, $p=.014$", then the given *p*-value is correct. Making this change throughout the document and re-running *statcheck* reveals only one inconsistency, "$t(84)= 3.11$, $p= .007$", which should have $p=.003$.

Third, this interpretation is consistent with data gleaned from Figures 2, 3, and 5 for the three experiments. Given the between-subjects design of each experiment, the ANOVA and contrast tests can be reconstructed from the means and standard errors. The only missing information is the sample size for each condition. R analysis scripts at the Open Science Framework demonstrate that a nearly homogeneous distribution of samples across conditions produces contrast *t*-values pretty close to the test statistics in Heyman and Ariely (2004). Additional scripts searched for a better distribution of sample sizes, and found good matches for almost all test statistics; albeit for some rather unusual sample sizes (e.g., Experiment 1 has *n*'s ranging from 36 to 136 across conditions). Ultimately, the data gleaned from the graphs suggest that the reported *p*-values and inferences in Heyman and Ariely (2004) might be valid. Although it seems possible to largely recover the analyses in this particular case, full verification, and examination with other analyses, requires long-term storage of data at some place like the Open Science Framework.

Gregory Francis

Department of Psychological Sciences, Purdue University

gfrancis@purdue.edu

References

Bauer, P. J., & Ariely, D. (2021). Expression of concern: Effort for payment: A tale of two markets. *Psychological Science*, *32*(8), 1338–1339. https://doi.org/10.1177/09567976211035782

Francis, G. (2021). Data gleaning for Heyman and Ariely (2004). *Open Science Framework*. https://osf.io/ns8zk/

Heyman, J., & Ariely, D. (2004). Effort for payment: A tale of two markets. *Psychological Science*, *15*(11), 787–793. https://doi.org/10.1111/j.0956-7976.2004.00757.x

Rife, S. C., Nuijten, M. B., & Epskamp, S. (2016). *statcheck: Extract statistics from articles and recompute p-values* [Web application]. http://statcheck.io