Before Data Analysis: Additional Recommendations for Designing Experiments to Learn about the World[1]

Andrew Gelman[2]

29 May 2023

Wedel and Gal (2023) make several recommendations for data analysis and reporting in psychology experiments: (1) summarize evidence in a continuous way, (2) recognize that rejection of statistical model A should not be taken as evidence in favor of preferred alternative B, (3) use substantive theory to generalize from experimental data to the real world, (4) report all the data rather than choosing a single summary, (5) report all steps of data collection and analysis.

All of these are good suggestions. We can go further by considering what comes before data analysis: design of experiments and data collection. The recommendations we give here are not new; our contribution here is just to put them in a convenient place and to remind experimenters that, once the data have been collected, the most important decisions have already been done.

Recommendation 1. Consider measurements that address the underlying construct of interest.

The concepts of validity and reliability of measurement are well known in psychology but are often forgotten in experimental design and analysis. Often we see exposure or treatment measures and outcome measures that connect only indirectly to substantive research goals. This can be seen in the frequent disconnect between the title and abstract of a research paper, on one hand, and the actual experiment, on the other. A notorious example in psychology is Hasan et al. (2013), who referred in their title to a "long-term experimental study" that in fact was conducted for only three days.

Our recommendation goes in two directions. First, set up your design and data collection to measure what you want to learn about. If you are interested in long-term effects, conduct a long-term study if possible. Second, to the extent that it is not possible to take measurements that align with your inferential goals, be open about this gap and explicit about the theoretical assumptions or external information that you are using to support your more general conclusions.

Recommendation 2. When designing an experiment, consider realistic effect sizes.

There is a tendency to overestimate effect sizes when designing a study. Part of this is optimism and availability bias: it is natural for researchers who have thought hard about a particular effect to think that it will be important, to envision scenarios where the

[1] To be published in *Journal of Consumer Psychology.* We thank Joel Huber for inviting this article and the U.S. Office of Naval Research for support of this work.

[2] Department of Statistics, Columbia University, New York.

treatment will have large effects and not to think so much about cases where it will have no effect or where it will be overwhelmed by other possible influences on the outcome. In addition, past results will be much more likely to be published if they have reached a significance threshold, and this results in literature reviews that vastly overestimate effect sizes.

Overestimation of effect sizes leads to overconfidence in design, with researchers being satisfied by small sample size and sloppy measurements in a mistaken belief that the underlying effect is so large that it can be detected even with a very crude inference. And this causes three problems. First, it is a waste of resources to conduct an experiment that is so noisy that there is essentially no chance of learning anything useful, and this sort of work can crowd out the more careful sorts of studies that would be needed to detect realistic effect sizes. Second, a false expectation of high power creates a cycle of hype and disappointment that can discredit a field of research. Third, the process of overestimation can be self-perpetuating, with a noisy experiment being analyzed until apparently statistically-significant results appear, leading to another overestimate to add to the literature. These problems arise not just in statistical power analysis (where the goal is to design an experiment with a high probability of yielding a statistically significant result) but also applies to more general design analyses where inferences will be summarized by estimates and uncertainties (Gelman, 2018).

Recommendation 3. Simulate your data collection and analysis on the computer first.

In the past, we have designed experiments and gathered data on the hope that the results would lead to insight and possible publication—but then the actual data would end up too noisy, and we would realize in retrospect that our study never really had a chance of answering the questions we wanted to ask. Such an experience is not a complete waste—we learn from our mistakes and can use them to design future studies—but we can often do better by preceding any data collection with a computer simulation.

Simulating a study can be more challenging than conducting a traditional power analysis. The simulation does not require any mathematical calculations; the challenge is the need to specify all aspects of the new study. For example, if the analysis will use regression on pre-treatment predictors, these must be simulated too, and the simulated model for the outcome should include the possibility of interactions.

Beyond the obvious benefit of revealing designs that look to be too noisy to detect main effects or interactions of interest, the construction of the simulation focuses our ideas by forcing us to make hard choices in assuming the structure and sizes of effects. In the simulation we can make assumptions about variation in measurement and in treatment effects, which can facilitate the first two recommendations above.

Recommendation 4. Design in light of analysis. Logician Raymond Smullyan (1979) wrote, “To know the past, one must first know the future.” The application of this principle to statistics is that design and data collection should be aligned with how you plan to analyze your data. As Sechest (2005) puts it, "The central issue is the validity of the inferences about the construct rather than the validity of the measure per se."

One place this arises is in the collection of pre-treatment variables. If there is concern about imbalance between treatment and control groups in an observational study or an experiment with dropout, it is a good idea to think about such problems ahead of time and gather information on the participants to use in post-data adjustments. Along similar lines, it can make sense to recruit a broad range of participants and record information on them to facilitate generalizations from the data to larger populations of interest. A model to address problems of representativeness should include treatment interactions so that effects can vary by characteristics of the person and scenario.

In summary, we can most effectively learn from experiment if we think plan the design and data collection ahead of time, which involves: (1) using measurement that relates well to underlying constructs of interest, (2) considering realistic effect sizes and variation, (3) simulating experiments on the computer before collecting any data, and (4) keeping analysis plans in mind in the design stage.

References

Andrew Gelman (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* 44, 16-23.

Youssef Hasan, Laurent Bègue, Michael Scharkow, and Brad J. Bushman (2013). The more you play, the more aggressive you become: A long-term experimental study of cumulative violent video game effects on hostile expectations and aggressive behavior. *Journal of Experimental Social Psychology* 49, 224-227.

Lee Sechrest (2005). Validity of measures is no simple matter. *Health Services Research* 40, 1584-1604.

Raymond Smullyan (1979). The chess mysteries of Sherlock Holmes. Knopf.

Michel Wedel and David Gal (2023). Beyond statistical significance: Five principles for the new era of data analysis and reporting. *Journal of Consumer Psychology.*