Tools for the Evaluation of the Quality of Experimental Research

René Bekkers, Center for Philanthropic Studies, VU Amsterdam[1], r.bekkers@vu.nl

November 11, 2016

Experiments can have important advantages above other research designs. The most important advantage of experiments concerns internal validity. Random assignment to treatment reduces the attribution problem and increases the possibilities for causal inference. An additional advantage is that control over participants reduces heterogeneity of treatment effects observed.

The extent to which these advantages are realized in the data depends on the design and execution of the experiment. Experiments have a higher quality if the sample size is larger, the theoretical concepts are more reliably measured, and have a higher validity. The sufficiency of the sample size can be checked with a power analysis. For most effect sizes in the social sciences, which are small (d = 0.2), a sample of 1300 participants is required to detect it at conventional significance levels ($p < .05$) and 95% power. Also for a stronger effect size (0.4) more than 300 participants are required.[2] The reliability of normative scale measures can be judged with Cronbach's alpha. A rule of thumb for unidimensional scales is that alpha should be at least .63 for a scale consisting of 4 items, .68 for 5 items, .72 for 6 items, .75 for 7 items, and so on.[3] The validity of measures should be justified theoretically and can be checked with a manipulation check, which should reveal a sizeable and significant association with the treatment variables.

The advantages of experiments are reduced if assignment to treatment is non-random and treatment effects are confounded. In addition, a variety of other problems may endanger internal validity.[4]

Also it should be noted that experiments can have important disadvantages. The most important disadvantage is that the external validity of the findings is limited to the participants in the setting in

---

[1] I wrote the following memo upon request of the editors of Nonprofit & Voluntary Sector Quarterly (NVSQ), the journal of the Association for Research on Nonprofit and Voluntary Action (ARNOVA), to assist them in the evaluation of the quality of research reporting on experiments. Views expressed here and errors in statements of fact are mine. I thank Rich Steinberg, Ivar Vermeulen, Camiel Beukeboom, Guido van Koningsbruggen, Mark van Vugt, Josh Tybur, Ashley Whillans and Artur Nilsson for comments and suggestions.

[2] Sample size calculations based on a t-test for equality of means. See table 1 in the appendix for criterion values at other effect sizes and other statistical tests.

[3] This rule of thumb is based on the formula for coefficient alpha assuming unidimensionality and an average interitem correlation of .30. Below .30 the construct is weak or multidimensional. If the average interitem correlation is higher, alpha increases. See Figure 1 in the appendix for a graphic illustration of criterion values at other values.

[4] Shadish, Cook & Campbell (2002) provide a useful list of such problems.

which their behavior was observed. This disadvantage can be avoided by creating more realistic decision situations, for instance in natural field experiments, and by recruiting (non-'WEIRD') samples of participants that are more representative of the target population. [5]

Recently, experimental research paradigms have received fierce criticism. Results of research often cannot be reproduced (Open Science Collaboration, 2015), publication bias is ubiquitous (Ioannidis, 2005). It has become clear that there is a lot of undisclosed flexibility, in all phases of the empirical cycle. While these problems have been discussed widely in communities of researchers conducting experiments, they are by no means limited to one particular methodology or mode of data collection. It is likely that they also occur in communities of researchers using survey or interview data.

In the positivist paradigm that dominates experimental research, the empirical cycle starts with the formulation of a research question. To answer the question, hypotheses are formulated based on established theories and previous research findings. Then the research is designed, data are collected, a predetermined analysis plan is executed, results are interpreted, the research report is written and submitted for peer review. After the usual round(s) of revisions, the findings are incorporated in the body of knowledge.

The validity and reliability of results from experiments can be compromised in two ways. The first is by juggling with the order of phases in the empirical cycle. Researchers can decide to amend their research questions and hypotheses after they have seen the results of their analyses. Kerr (1989) labeled the practice of reformulating hypotheses HARKING: Hypothesizing After Results are Known. Amending hypotheses is not a problem when the goal of the research is to develop theories to be tested later, as in grounded theory or exploratory analyses (e.g., data mining). But in hypothesis-testing research harking is a problem, because it increases the likelihood of publishing false positives. Chance findings are interpreted *post hoc* as confirmations of hypotheses that *a priori* are rather unlikely to be true. When these findings are published, they are unlikely to be reproducible by other researchers, creating research waste, and worse, reducing the reliability of published knowledge.

The second way the validity and reliability of results from experiments can be compromised is by misconduct and sloppy science within various stages of the empirical cycle (Simmons, Nelson & Simonsohn, 2011). The data collection and analysis phase as well as the reporting phase are most vulnerable to distortion by fraud, p-hacking and other questionable research practices (QRPs).

- In the data collection phase, observations that (if kept) would lead to undesired conclusions or non-significant results can be altered or omitted. Also, fake observations can be added (fabricated).

[5] As Henrich, Heine & Norenzayan (2010) noted, results based on samples of participants in Western, Educated, Industrialized, Rich and Democratic (WEIRD) countries have limited validity in the discovery of universal laws of human cognition, emotion or behavior.

- In the analysis of data researchers can try alternative specifications of the variables, scale constructions, and regression models, searching for those that 'work' and choosing those that reach the desired conclusion.
- In the reporting phase, things go wrong when the search for alternative specifications and the sensitivity of the results with respect to decisions in the data analysis phase is not disclosed.
- In the peer review process, there can be pressure from editors and reviewers to cut reports of non-significant results, or to collect additional data supporting the hypotheses and the significant results reported in the literature.

Results from these forms of QRPs are that null-findings are less likely to be published, and that published research is biased towards positive findings, confirming the hypotheses, published findings are not reproducible, and when a replication attempt is made, published findings are found to be less significant, less often positive, and of a lower effect size (Open Science Collaboration, 2015).

Alarm bells, red flags and other warning signs

Some of the forms of misconduct mentioned above are very difficult to detect for reviewers and editors. When observations are fabricated or omitted from the analysis, only inside information, very sophisticated data detectives and stupidity of the authors can help us. Also many other forms of misconduct are difficult to prove. While smoking guns are rare, we can look for clues. I have developed a checklist of warning signs that editors and reviewers can use to screen submissions (see appendix). The checklist uses terminology that is not specific to experiments, but applies to all forms of data. While a high number of warning signs in itself does not prove anything, it should alert reviewers and editors. There is no norm for the number of flags. Those who would like to count good practices and reward authors for a higher number can count gold stars rather than red flags. The checklist was developed independently of the checklist that Wicherts et al. (2016) recently published.

With the increasing number of retractions of articles reporting on experimental research published in scholarly journals the awareness of the fallibility of peer review as a quality control mechanism has increased. Communities of researchers employing experimental designs have formulated solutions to these problems. In the review and publication stage, the following solutions have been proposed.

- Access to data and code. An increasing number of science funders require grantees to provide open access to the data and the code that they have collected. Likewise, authors are required to provide access to data and code at a growing number of journals, such as Science, Nature, and the American Journal of Political Science. Platforms such as Dataverse, the Open Science Framework and Github facilitate sharing of data and code. Some journals do not require access to data and code, but provide Open Science badges for articles that do provide access.

- Pledges, such as the '21 word solution', a statement designed by Simmons, Nelson and Simonsohn (2012) that authors can include in their paper to ensure they have not fudged the data: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study."
- Full disclosure of methodological details of research submitted for publication, for instance through psychdisclosure.org is now required by major journals in psychology.
- Apps such as Statcheck, p-curve, p-checker, and r-index can help editors and reviewers detect fishy business. They also have the potential to improve research hygiene when researchers start using these apps to check their own work before they submit it for review.

As these solutions become more commonly used we should see the quality of research go up. The number of red flags in research should decrease and the number of gold stars should increase. This requires not only that reviewers and editors use the checklist, but most importantly, that also researchers themselves use it.

The solutions above should be supplemented by better research practices before researchers submit their papers for review. In particular, two measures are worth mentioning:

- Preregistration of research, for instance on aspredicted.org. An increasing number of journals in psychology require research to be preregistered. Some journals guarantee publication of research regardless of its results after a round of peer review of the research design.
- Increasing the statistical power of research is one of the most promising strategies to increase the quality of experimental research (Bakker, Van Dijk & Wicherts, 2012). In many fields and for many decades, published research has been underpowered, using samples of participants that are not large enough the reported effect sizes. Using larger samples reduces the likelihood of both false positives as well as false negatives.

A variety of institutional designs have been proposed to encourage the use of the solutions mentioned above, including reducing the incentives in careers of researchers and hiring and promotion decisions for using questionable research practices, rewarding researchers for good conduct through badges, the adoption of voluntary codes of conduct, and socialization of students and senior staff through teaching and workshops. Research funders, journals, editors, authors, reviewers, universities, senior researchers and students all have a responsibility in these developments.

References

Bakker, M., Van Dijk, A. & Wicherts, J. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, 7(6): 543–554.

Henrich, J., Heine, S.J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33: 61 - 135.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. PLoS Medicine, 2(8): e124. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

Kerr, N.L. (1989). HARKing: Hypothesizing After Results are Known. *Personality and Social Psychology Review*, 2: 196-217.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science*, 349. http://www.sciencemag.org/content/349/6251/aac4716.full.html

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22: 1359–1366.

Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2012). A 21 Word Solution. Available at SSRN: http://ssrn.com/abstract=2160588

Wicherts, J.M., Veldkamp, C.L., Augusteijn, H.E., Bakker, M., Van Aert, R.C & Van Assen, M.L.A.M. (2016). Researcher degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers of Psychology, 7*: 1832. http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01832/abstract

Appendix

Table 1. Criterion values for sample sizes in t-tests for difference of means

Effect size	Critical t	one-tailed	Critical t	two-tailed
.50	1.654	176	1.971	210
.40	1.650	272	1.967	328
.30	1.648	484	1.964	580
.20	1.646	1084	1.962	1302

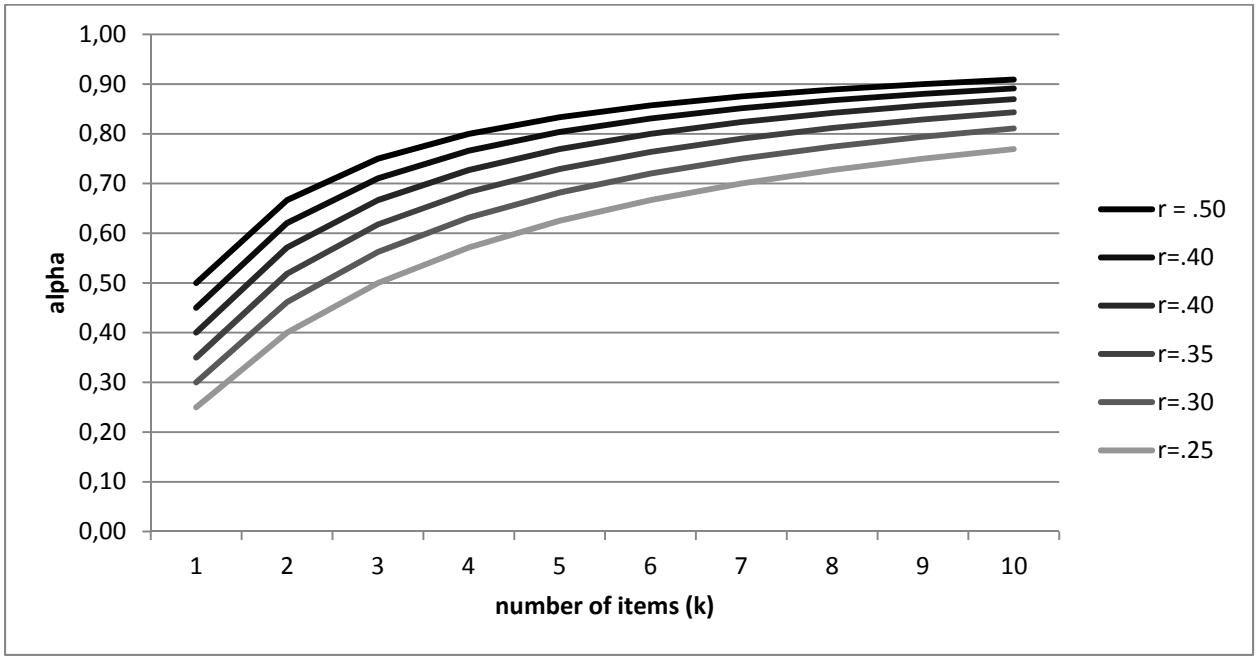
Sample size required for significant t-test statistics for the difference between independent means of two groups of equal size, at α = .05 (power =95%), computed with G*Power 3.1.9.2

Table 2. Criterion values for sample sizes in exact tests for correlations

Effect size	Critical r	one-tailed	Critical r	two-tailed
.50	.27	38	.29	46
.40	.21	63	.23	75
.30	.15	115	.17	138
.20	.10	266	.11	319

Sample size required for significant bivariate normal correlations, at α = .05 (power =95%), computed with G*Power 3.1.9.2

Figure 1. Criterion values for coefficient alpha, * Calculated with $\alpha = k * r / (1 + r(k-1))$



The figure displays values for coefficient alpha given the number of items (k) and the average interitem correlation between scale items (r). You can use the graph to determine the interitem correlation when α and k are reported. For instance, if $k = 6$ and $\alpha = .60$, $r = .25$.

RESEARCH QUALITY CHECKLIST

Warning signs		Real results	
The power of the analysis is too low.		The power of the analysis is high.	
The results are too good to be true.		The results are not perfect, noisy.	
All hypotheses are confirmed.		Some hypotheses are rejected.	
P-values are just below critical thresholds (e.g., p<.05)		P-values vary throughout the paper regardless of thresholds.	
A groundbreaking result is reported but not replicated in another sample.		The groundbreaking result is replicated in a subsequent experiment, preferably elsewhere among a larger sample.	
The data and code are not made available upon request.		The data and code are made available to reviewers upon request.	
The data are not made available upon article submission.		The data are made available to reviewers and future readers upon article submission.	
The code is not made available upon article submission.		The code is made available to reviewers and future readers upon article submission.	
Materials (manipulations, survey questions) are described superficially.		All the materials (manipulations, survey questions) are made available to reviewers.	
Descriptive statistics are not reported.		Descriptive statistics are reported.	
The hypotheses are tested in analyses with covariates and results without covariates are not disclosed.		Analyses without covariates are reported.	
The research is not preregistered.		The research is preregistered.	
No details of an IRB procedure are given.		Details of the IRB procedure (time, location, file number) are given.	
Participant recruitment procedures are not described.		Participant recruitment procedures are described.	
Exact details of time and location of the data collection are not described.		Time(s) and location(s) of the data collection are fully described.	
A power analysis is lacking.		A power analysis is provided to justify the sample size and the effect size is credible.	
Unusual / non-validated measures are used without justification.		Well-known and validated measures are used.	
Different dependent variables are analyzed in different studies within the same article without justification.		The same dependent variables are analyzed in different studies.	
Variables are (log)transformed or recoded in unusual categories without justification.		Variables are recoded in conventional categories or (log)transformed with theoretical justification.	
Numbers of observations mentioned at different places in the article are inconsistent. Loss or addition of observations is not justified.		Numbers of observations mentioned at different places are consistent, and if not, changes of numbers of observations are justified.	
A one-sided test is reported when a two-sided test would be appropriate.		Conservative significance tests are reported.	
Test-statistics (p-values, F-values) reported are incorrect.		Test-statistics reported are correct.	
NUMBER OF FLAGS:		*NUMBER OF STARS:*	

General quality indicators

In addition to the flags that may indicate sloppy science and misconduct, there are general indicators of quality for experiments. A higher number of problems should alert editors and reviewers. Conversely, authors could be rewarded with a higher number of stars when their research includes a greater number of positive characteristics.[6]

Worse		Better	
The number of observations is too small to detect the assumed effect size.		The number of observations is sufficiently large to detect the assumed effect size.	
Experimenters were aware of the hypotheses.		Experimenters were blind to the hypotheses.	
Participants were aware of the hypotheses.		Participants were blind to the hypotheses.	
Reliability coefficients are not reported.		Reliability coefficients are reported and sufficiently high given the number of items.	
Manipulation checks are lacking.		Manipulation checks are reported.	
The intended treatment effects are confounded.		The treatment is clean, limited to the theoretically relevant effect.	
The participants are a selective sample of the group to which the findings are generalized.		The participants are representative for the group to which the findings are generalized.	
NUMBER OF FLAGS:		*NUMBER OF STARS:*	

[6] This is a partial list. A complete list adapted from Shadish, Cook & Campbell (2002) is posted at http://csrakes.yolasite.com/resources/Presentations/Threats_to_Validity.pdf