Asanov, Igor; Bühren, Christoph; Zacharodimou, Panagiota

Working Paper

The power of experiments: How big is your n?

MAGKS Joint Discussion Paper Series in Economics, No. 32-2020

Provided in Cooperation with:
Faculty of Business Administration and Economics, University of Marburg

Suggested Citation: Asanov, Igor; Bühren, Christoph; Zacharodimou, Panagiota (2020) : The power of experiments: How big is your n?, MAGKS Joint Discussion Paper Series in Economics, No. 32-2020, Philipps-University Marburg, School of Business and Economics, Marburg

This Version is available at:
https://hdl.handle.net/10419/234837

Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 32-2020

Igor Asanov, Christoph Bühren and Panagiota Zacharodimou

The Power of Experiments: How Big is your *n*?

This paper can be downloaded from
http://www.uni-marburg.de/fb02/makro/forschung/magkspapers

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

The power of experiments: How big is your *n*?

Igor Asanov[1], Christoph Bühren[2], and Panagiota Zacharodimou[3]

Abstract

The replicability and credibility crisis in psychology and economics sparked the debate on underpowered experiments, publication biases, and p-hacking. Analyzing the number of independent observations of experiments published in Experimental Economics, Games and Economic Behavior, and the Journal of Economic Behavior and Organization, we observe that we did not learn much from this debate. The median experiment in our sample has too few independent observations and, thus, is underpowered. Moreover, we find indications for biases in reporting highly significant results. We investigate for which papers and experiments it is more likely to find reporting biases, and we suggest remedies that could help to overcome the replicability crisis.

Keywords: Statistical power; statistical significance; meta-study; balanced randomization; caliper test

JEL: C10; C12; C18

[1] University of Kassel, Department of Economics, igor.asanov@uni-kassel.de

[2] Corresponding author: Clausthal University of Technology, Department of Economics, Julius-Albert Str. 2, 38678 Clausthal-Zellerfeld, Germany, christoph.buehren@tu-clausthal.de

[3] European Parliament, zacharodimou.p@gmail.com

1. Introduction

In the last decade, researchers, reviewers, and editors in experimental economics demanded to focus more on the power of experiments. Underpowered studies (i.e., studies with a power lower than 80%) are less likely to find true effects. More severe and less obvious, underpowered studies inflate estimates undermining the purpose of scientific research (Button et al., 2013; Cohen, 1988). Publication decisions just on basis of statistically (and not economically) significant results and the aim of researchers to obtain statistically significant results lead to publication bias (Rosenthal, 1979), reporting biases, such as p-hacking (Simonsohn, Nelson, and Simmons, 2014), and HARKing (Hypothesizing After the Results are Known, Kerr, 1998). As a consequence, we likely have a biased view of (economic) results based on published papers.

Unfortunately, replication studies in economics are very rare – 0.1% according to Mueller-Langer et al. (2019). It is little wonder that replications often fail to find the originally published effect sizes. Klein et al. (2014) and Klein et al. (2018), e.g., tried to replicate with their many labs project prominent experiments in psychology and behavioral economics. For anchoring, the authors find even higher effect sizes than in the published papers. In most of the other replications, however, the effect sizes were much lower than in the publications. This discrepancy was highest for priming experiments, for which the mean effect size in the many lab project was zero. The replication crisis in psychology, management, and economics led to a credibility crisis (see Anderson et al., 2015, Gilbert et al., 2016, Bergh et al., 2017, Byington and Felps, 2017, and Ioannidis, Stanley, and Doucouliagos, 2017).

Analyzing a random sample of experiments published in three prominent field journals of experimental economics from 2011 to 2017, we show that the vast majority of experiments have by far too few independent observations and, thus, are underpowered in particular if experiments employ session-level randomization. Moreover, we find reporting biases especially for highly significant results in papers without hypotheses. The remainder of our paper is structured as follows: The next section reviews related literature on publication biases, p-hacking, and power of experiments. Section 3 derives our hypotheses, followed by a description of the data and methods that we use (Section 4). Section 5 reports our results, and the last section concludes and discusses our results.

2. Related Literature

Ioannidi's (2005) question "Why most published research findings are false?" started a discussion on the reliability and replicability of empirical findings in many disciplines (e.g., Pashler and Harris, 2012; Baker, 2016, Goodman, Fanelli, and Ioannidis (2016); Munafò et al., 2017; Benjamin et al., 2018; Camerer et al., 2018), including economics (e.g., Dreber et al., 2015; Camerer et al., 2016; Berry et al., 2017; Butera and List, 2017; Coffman, Niederle and Wilson, 2017; Maniadis et al., 2017; Drazen et al., 2019; Butera et al., 2020). Fanelli, Costas, and Ioannidis (2017) argue that scientific research across all disciplines should satisfy a minimum of standards to overcome problematic practices. The two most severe problems are underpowered studies and reporting biases. Low statistical power results in a low probability of finding effects that are true (Ioannidis, 2005). The tendency to deifying statistical significant results induces biases: cherry-picking of i) studies that yield statistically significant results – publication bias or 'file drawer problem' (Rosenthal, 1979), ii) the statistically significant results of a study (Simonsohn, Nelson and Simmons, 2014), or iii) the best hypotheses that support statistically significant results post hoc (Kerr, 1998).

Fisher (1925) introduced the idea of significance levels and proposed the p-value as a tool to examine whether the sample provides evidence against the null hypothesis. (Fisher, 1932: p.82) proposed the cut-off point $p = 0.05$:

"*If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05.*"

Since then, statistical significance at the 5%-level determined to a considerable extent scientific decisions in fields such as psychology, economics, medicine, and biology (e.g., Ziliak and McCloskey, 2008; Andrews and Kasy, 2019). For the academic community in economics, the 10%-level has been acceptable as well (Brodeur et al., 2016), whereas highly significant results are typically associated with p-values below 1%. Statistical significance was converted from a researcher's tool to the researcher's aim. As a consequence, reporting biases became more likely (Rosenthal, 1979; Simonsohn, Nelson, and Simmons, 2014).

Should the publication decision depend on whether p-values of a study lie below a threshold, studies that provide significant results ($p < 0.05$) are more likely to get published, whilst those that provide non-significant results ($p > 0.05$) are more likely to be placed into the file drawer, and never be seen and read by others (Rosenthal, 1979). This publication bias

encompasses the dissemination of a significant amount of studies that provide only false-positive results and the withholding of studies that provide non-significant results, which can still be important and provide substantial information if assessed and discussed (Abadie, 2020). Publication biases can happen either before the time of submission – when authors decide not to submit studies with non-significant results for publication – or in the reviewing process when reviewers reject studies with non-significant results (Rosenthal, 1979; Franco, Malhotra and Simonovits, 2014). The most severe case of the file drawer problem would be that the 5% of the studies that produce false-positive findings are published in journals, and the remaining 95% of the studies that produce negative results are unpublished and hidden in researchers' and reviewers' file drawers (Rosenthal, 1979). As a consequence, published results are possibly an inaccurate portrayal of reality (Ioannidis, 2008; Pashler and Harris, 2012).

Instead of not submitting the whole study for publication, researchers could report solely the findings that meet the criterion of statistical significance or manipulate and analyze the data in such a way that produces significant results (Simonsohn, Nelson, and Simmons, 2014). In particular, 'p-hacking' includes reporting only the measures and conditions that pass a certain threshold for statistical significance, excluding participants, using covariates, transforming the data, and adjusting the sample size (Simonsohn, Nelson, and Simmons, 2014). Synonyms of p-hacking are ‘data snooping’ (White, 2000), ‘bias’ (Ioannidis, 2005), ‘significance chasing’ (Ioannidis and Trikalinos, 2007), and ‘specification searches’ (Gerber and Malhotra, 2008). A further reporting bias is to formulate a hypothesis that supports a significant finding post hoc, but presenting it as an a priori hypothesis - ‘HARKing’ (Kerr, 1998). A similar dimension of HARKing is concealing an a priori hypothesis that provided negative findings (Kerr, 1998). Practices of selective reporting of statistically significant results increase the chances of a study to be published. Nonetheless, researchers might not deliberately try to deceive reviewers, readers, and the academic community (Simmons, Nelson, and Simonsohn, 2011; Simonsohn, Nelson, and Simmons, 2014). They could tend to subconsciously process ambiguous information in a biased manner to favor the conclusion they want to reach (Kunda, 1990; Bastardi, Uhlmann, and Ross, 2011; Simmons, Nelson, and Simonsohn, 2011).

Statistical significance is, by definition, a tool for null hypothesis testing rather than a tool for statistical inference. However, many researchers confuse statistically significant with economically significant findings: Ziliak and McCloskey (2004) review articles published in American Economic Review in the 1990s and find that of the studies including at least one test of statistical significance, 82% identified statistical significance with economic significance. In

this regard, there is an increasing concern in the academic community that in the pursuit of the longed-for statistical significance, p-values could be misused and misinterpreted (Goodman, 1999 and 2008; Hubbard and Lindsay, 2008; Ziliak and McCloskey, 2008; Biau, Jolles and Porcher, 2010; Greenland et al., 2016; Vidgen and Yasseri, 2016). Wasserstein and Lazar (2016) highlight that the p-value by itself provides poor evidence against or in favor of the null hypothesis, and that it should not be considered alone, but together with other statistical measures. Statistical power, e.g., should be used to evaluate the importance of a scientific result (Ziliak and McCloskey, 2008). It reveals the probability of finding an effect given this effect is true, and it is inversely related to the false-negative rate. In other words, low statistical power signifies more false-negative inferences and a lower chance of discovering a genuine effect. For example, designing an experiment with a power of 80% implies expecting to find 80 of 100 true effects of a given magnitude and significance level. Experiments with a power of 20% will only expect to find 20 of the 100 true effects of a given magnitude and significance level. More severe, an underpowered experiment is accompanied by a low probability that a significant research finding is true (Ioannidis, 2005; Moonesinghe, Khoury, and Janssens, 2007; Button et al., 2013; Munafò et al., 2017; Czibor, Jimenez-Gomez and List, 2019).

However, Ziliak and McCloskey (2004) observe that only 8% of the articles published in the American Economic Review in the 1990s considered the power of the tests – in the 1980's it was only 4.4% (McCloskey and Ziliak,1996). Ioannidis, Stanley, and Doucouliagos (2017) calculate a median power of 18% in a meta-analysis of 6,700 empirical economic studies. Zhang and Ortmann (2013) review 95 articles from Experimental Economics between 2010 and 2012, none of which considered optimal sample size and statistical power. In a meta-analysis of dictator games, Zhang and Ortmann (2013) calculate a median power of 25%. Bellemare, Bissonnette, and Kröger (2014) examine experimental practices in 116 experiments published in Experimental Economics between 2012 and 2013 and conclude that statistical power did not play any role in determining the sample sizes of the experiments. DellaVigna and Linos (2020), studying two well-powered Nudge Unit trials in the United States and another sample of nudge trials published in academic journals from two meta-analyses, found that in the Nudge Unit trials the average impact of nudge interventions is highly statistically significant but smaller at 1.4 percentage points. They showed that this difference could be explained to a great extent by publication bias and low statistical power.

3. Hypotheses

The nature of our study is descriptive and explorative. Nevertheless, we formulated hypotheses before collecting and analyzing our data. First, given the previous literature (e.g., Zhang and Ortmann, 2013), we expect that the laboratory experiments are underpowered.

Hypothesis 1a: The median laboratory experiment does not achieve the minimum desired level of statistical power.

Our main focus is to consider the unit of randomization – in particular, individual or session-level randomization – and, thus, assess the power of the experiments from this perspective. In case of a higher level of randomization – session – the power can be drastically decreased compared to a lower level of randomization – individual – even though one would have the same number of participants. This issue is often referred to as the "design effect" (Mead, Gilmour, and Mead, 2012; Glennerster and Takavarasha, 2013). Unless we observe that the design effect is taken into account by researchers and the experiments with a higher level of randomization have a larger number of participants, we expect that experiments with session-level of randomization will be even more underpowered compared to experiments that randomize at the individual level.

Hypothesis 1b: The median laboratory experiment that randomizes at the session-level has lower statistical power than the median experiment that randomizes at the individual level.

Brodeur et al. (2016) analyzed the distribution of z-statistics for laboratory experiments and found that discontinuities around significance thresholds were less visible compared to other samples. However, we hypothesize to find evidence of reporting biases just below the three conventional levels of statistical significance.

Hypothesis 2a: The incidence of reporting biases is present around the 0.01 significance level.
Hypothesis 2b: The incidence of reporting biases is present around the 0.05 significance level.
Hypothesis 2c: The incidence of reporting biases is present around the 0.1 significance level.

Next, we turn to the analysis of the sub-sample heterogeneity to provide explorative analysis of the heterogeneity. We presume that the occurrence and strength of reporting biases may be associated with article characteristics, reporting techniques, and experimental practices. Considering that the academic community has progressively recognized the problem of reporting biases, we hypothesize that the adoption of practices such as *p*-hacking or HARKing declines over time.

Hypothesis 3a: The incidence of reporting biases decreases over time.

The occurrence and intensity of reporting biases may vary with the number of authors. Collaboration, as well as the combination of skills and knowledge, could translate to less subjective scientific work (Hudson, 1996). Thus, we believe:

Hypothesis 3b: The higher the number of authors that are involved with the research and the article, the lower the levels of reporting biases.

Moreover, without speculating on a ranking of Experimental Economics, Games and Economic Behavior, and Journal of Economic Behavior and Organization, we think:

Hypothesis 3c: The incidence of reporting biases varies with the journal.

We further suspect that there is a link between the occurrence of reporting biases and the presentation of empirical results. Results that are statistically significant and/or support the alternative hypotheses of the article are usually reported as the main results. Therefore, reporting of tests could be more biased if they belong to main results rather than robustness checks (Brodeur et al., 2016).

Hypothesis 3d: The incidence of reporting biases is higher among the main results than among non-main results.

Furthermore, Brodeur et al. (2016) find that reporting biases are more prevalent among tests that use eye-catchers to denote statistical significance, which leads us to:

Hypothesis 3e: The incidence of reporting biases is higher among tests denoting statistical significance merely through eye-catchers than among tests explicitly declaring p-values.

The next two hypotheses are also based on Brodeur et al. (2016), who argue that reporting biases can be restricted if the experimental paper relies on a list of hypotheses and/or a formal model.

Hypothesis 3f: Articles without an explicit list of hypotheses are more vulnerable to biased reporting.

Hypothesis 3g: Articles without a theoretical model are more vulnerable to biased reporting.

Compared to one-shot experiments, experiments that are repeated in more rounds offer a higher number of observations per subjects (but the same number of independent observations).

Renkewitz, Fuchs, and Fiedler (2011) observe stronger discontinuities in p-values of experiments with more rounds compared to one-shot experiments. Thus, we hypothesize:

Hypothesis 3h: The incidence of reporting biases is higher if the experiment uses repeated games than one-shot games.

Similarly, reporting biases are likely to vary also by the number of independent observations:

Hypothesis 3i: The incidence of reporting biases is influenced by the number of independent observations.

4. Data and Methods

We randomly drew articles published in Experimental Economics, Games and Economic Behavior, and Journal of Economic Behavior and Organization between 2011 and 2017. For each journal and each year we randomly sampled 10 articles by generating random numbers corresponding to the issues in the volumes, and articles.[4] Thus, we draw a random sample of 182 articles in those three journals with 10 articles per journal per year. 78 articles presented laboratory experiments, but 9 of them do not provide any empirical tests in tables in the paper. Thus, we provide an analysis of 69 articles which included at least one empirical test of laboratory experiment provided in a table (the appendix lists all these papers). We followed Brodeur et al. (2016) and Bruns et al. (2019) in collecting general and statistical information of the articles and included specific information on the experiments. We use exactly the data reported in the original papers without rounding and included all treatment and control variables for which tests are reported in the paper – excluding descriptive statistics. Table 1 lists the paper- and test-specific variables that we collected.

[4] We used a web tool that generates random numbers, namely random.org (Haahr, 2019). For each generation of random numbers, a specific timestamp was automatically generated (e.g., Timestamp: 2019-03-21 20:29:11 UTC). For transparency purposes, all the random numbers generated and their timestamps were recorded and kept.

Table 1. Included variables

Article-specific variables	
Year	Year of publication
Journal	Name of the journal
Volume	Volume of the journal
Issue	Issue of the journal
Title	Title of the article
N.authors	Number of authors
Author_#	First and last name(s) of author #
Sample.size	Sample size of the laboratory experiment
Treatments	Number of the treatments in the laboratory experiment
Formal.model	1 = formal model, 0 = otherwise
List.hypotheses	1 = explicit list of hypotheses, 0 = otherwise
One.shot	1 = one-shot experiment, 0 = otherwise
Test-specific variables	
N.pvalue	Running number for each test per article
Eye.catchers	1 = statistical significance reported through eye-catchers, 0 = otherwise
Between.subj	1 = between-subject design, 0 = otherwise
Within.subj	1 = within-subjects design, 0 = otherwise
Between.sess	1 = between-session design, 0 = otherwise
Within.sess	1 = within-session design, 0 = otherwise
Balanced.observables	1 = randomization balanced on observables characteristics, 0 = otherwise
Regression	1 = regression, 0 = otherwise
Test.type	Statistical method or type of test
Controls.before	1 = controls before the treatment that can be affected by treatment (basically, any except demographic characteristics: age, gender, faculty), 0 = otherwise
Controls.after	1 = controls after the treatment that can be affected by treatment (basically, any except demographic characteristics: age, gender, faculty), 0 = otherwise
Controls.from	1 = controls from experiment for non-treatment variables (e.g. risk preferences, gender), 0= otherwise
Controls.question	1 = controls from end-of-session questionnaire for non-treatment variables (e.g. risk preferences, gender), 0 = otherwise
Indep.obs	Number of independent observations of the test
Subj	Number of subjects of the test
Obs	Number of observations of the test
Two.sided	1 = two-sided test, 0 = one-sided test
Appendix	1 = test comes from a model that belongs to an appendix, 0 = otherwise
Robustness	1 = test comes from a model that is explicitly declared to be a robustness check or extension, or the table belongs to a section that conducts robustness checks, 0 = otherwise

Pvalue	p-value of the test as reported in the article if available
Tstat	t-statistic of the test as reported in the article if available
Zstat	z-statistic of the test as reported in the article if available
Coef	coefficient of the test as reported in the article if available
Se	standard error of the test as reported in the article if available
P.eyecatcher	Reported significance levels by means of eye-catchers (0.1 = <0.1, 0.05 = <0.05, 0.01 = <0.01, 0.001 = <0.001, 0 = not significant)
Source	Table, row, and column from which the test is extracted

Only a small number of papers present more than one experiment – in these cases, the sample size, the number of sessions, and the number of treatments varies across the tests of the paper. In a few studies, the number of subjects is not explicitly mentioned but can be calculated. We consider if the paper includes a theoretical model and if the authors use an explicit list of hypotheses. Moreover, we control if the experiment is a one-shot game or consists of more rounds.

We assess for each test of the experiment whether it is based on a within- or between-subject design and check the level of randomization: between-session randomization if sessions are assigned to treatments or within-session randomization if individuals within a session are assigned to treatments. Additionally, we assess if the randomization is balanced on observables characteristics of the subjects like age or gender (Bruhn and McKenzie, 2009). We observe whether the tests are in regressions or non-parametric tests and whether they are one- or two-sided. Furthermore, we analyze if the tests employ eye-catchers, such as stars or letters, to signal statistical significance. We also distinguish between main results and robustness checks – in cases of ambiguities, we conservatively consider the test as a main result. For every test, we collect the test statistics, coefficients, standard errors, and p-values or the significance levels.

We take into account if the tests use control variables: i) variables that can be affected by treatment and are collected before the treatment, ii) variables that can be affected by treatment and are collected after the treatment, iii) non-treatment variables that are collected before the treatment, and iv) non-treatment variables that are collected after the treatment. Control variables that can be affected by treatment cannot be variables such as demographic characteristics, age, faculty, or gender. Conversely, non-treatment control variables can be gender, age, demographic characteristics, or personality traits such as risk aversion, other-regarding preferences, or the 'Big-Five' personality dimensions. Questionnaires that elicit this kind of information (i.e., non-treatment variables) can be placed at the beginning or at the end of a session. In case that subjects have to fill in a questionnaire containing questions to their personality characteristics and attitudes after completing the experimental tasks, it is important

to bear in mind that that the specific characteristics may be influenced by the experiment. In the z-tree software (Fischbacher, 2007), the questionnaires are run at the end of a session (Fischbacher, Bendrick, and Schmid, 2015). Given that most economic laboratory experiments are programmed with z-tree, it is reasonable to expect that questionnaires are conducted after the experiment. In the majority of the papers (46 out of 69 articles), the authors confirmed the use of z-tree. A few articles included non-treatment variables in statistical tests without explicitly stating whether these variables were extracted from pre-session or post-session questionnaires. If these experiments employed z-tree, we coded them as variables from a post-experimental questionnaire; if not, we code them as non-treatment variables from a pre-session questionnaire.

We collect the number of subjects in the experiment and the number of observations: the number of choices that subjects make during the experiment. Finally, we analyze the number of independent observations per test at the level of randomization, thus, aiming to assess the *effective sample size*: the sample size that accounts for the correlation between observations.

To illustrate the necessity to analyze independent observations or the effective sample size, consider the following example from educational economics. The researchers have access to 6,000 students in 6 schools to test the effect of an educational program on a knowledge test. Due to the practical reasons they randomly allocate 3 schools to the treatment group and 3 schools to the control group. Thus, researchers can treat up to 3,000 students and observe up to 3,000 students in the control condition. Now, they think about how many of them to treat given budget limitations. A large number of observations is better. However, schools differ, e.g. students from private schools in rich areas will probably get higher results on the knowledge test. Adding additional students not necessarily yield entirely new information, as values are not independent. To which extent it undermines the study depends on the level that differences in the knowledge test can be explained by school characteristics (cluster) – intra-cluster correlation. Formally, intra-cluster correlation (ICC) is defined by:

$$\rho = \frac{s_b^2}{s_b^2 + s_w^2}$$

, where s^2_b is the variance between clusters and s^2_w the variance within clusters. If $\rho = 0$, no variance is explained by cluster characteristics. If $\rho = 1$, all the variance is explained by cluster characteristics and the *effective sample size* is equal to the number of independent observations. In our example, if $\rho = 1$, the *effective sample size* is equal to 6 (i.e., the number of schools), which leads to an underpowered study.

In the context of economic laboratory experiments, if researchers randomize within-subject or within-session with the unit of randomization - subject, one can treat each subject as an independent observation. However, if researchers randomize between groups of subjects, one has to consider the group-specific effect. It is particularly pronounced when researchers randomize on the session level, where session-specific characteristics affect the data generating process: time of the session (morning/evening session, weekday, month, year of the session), time of invitation for the experiment, self-selection of subjects to specific session slots, differences in running the sessions (different research assistants, technical problems, unplanned deviations from study protocol), etc. As a consequence, one has to take non-independence of sessions into account and consider observations on the level of sessions as independent observations. Thus, we focus on the number of independent observations per test at the level of randomization. In within-subject designs, the number of independent observations is equal to the total number of subjects of all sessions. In between-subject designs, it is typically equal to the total number of sessions. There are very few cases in which experimenters use between-subject designs, but randomize within-session, that is run all the treatments in each session.

5. Results

Out of the total number of tests we collected, 71% are published in *Experimental Economics* 16% in the *Journal of Economic Behavior and Organization* and 13% in *Games and Economic Behavior*. The average number of tests per article is 46, the median is 40. The minimum number of tests per article is 3 and the maximum 185. The vast majority of the articles (89%) are co-authored – the median number of authors is two. Similarly, Card and Dellavigna (2013) observe an average of 2.3 authors of papers published in top economics journals in 2012. Less than half of the papers provide a theoretical model accompanied by hypotheses before the results section. The number of tests is slightly higher in these papers compared to papers without theoretical models and hypotheses.

Table 2. Descriptive statistics

	Number of…	
	Articles	**Tests**
Experimental Economics	49	2,264
	[71]	[71]
Economic Behavior and Organization	13	510
	[19]	[16]
Games and Economic Behavior	7	398
	[10]	[13]
Single-authored	8	350
	[12]	[11]
Double-authored	31	1,540
	[45]	[49]
Multiple-authored (authors ≥ 3)	30	1,282
	[43]	[40]
Formal model	33	1,729
	[48]	[55]
Explicit List of Hypotheses	34	1,751
	[49]	[55]
One-shot experiments	15	488
	[22]	[15]
Between-subject randomization		2,591
		[82]
Between-session randomization		2,510
		[79]
Regression		2,568
		[81]
Two-sided tests		3,046
		[96]
Randomization balanced on observables		0
		[0]
Main Results		2,699
		[85]
Using eye-catchers		2,655
		[84]

Notes: This table reports the number of tests and tests for each sub-category. Proportions relative to the total sample are mentioned in brackets.
Source: *Experimental Economics, Journal of Economic Behavior and Organization* and *Games and Economic Behavior*, 2011-2017. See the list of laboratory experiments included for the analysis in the appendix.

The sample size ranges from 30 to 895 subjects, and the median number of subjects is 180. The median number of treatments equals four. Around 80% of the experiments use a between-subject design and randomize between sessions – only two between-subject designs allocate more than one treatment to a session. 22% of the experiments are one-shot games. We classify 85% of the tests in our sample as the main results. Most of the studies use regression analyses and nearly all of the tests are two-sided. 84% of the tests use eye-catchers, such as stars, to indicate significance levels. Not a single paper in our sample balances the randomization on observable characteristics of the subjects across treatments.

A few tests control for non-treatment variables elicited from post-experimental questionnaires, such as gender and risk preferences. The average number of subjects per test is approximately 146, and the median 98. However, the average number of independent observations per test is around 53, and the median number of independent observations per test is only eight: 6 for tests in a between-subject design and 120 for tests in a within-subject design. Figure 1 illustrates the discrepancy of subjects (upper panel) vs. independent observations (lower panel) in between-subject designs (left panel) – for within-subject designs, subjects and independent observations are identical (right panel).

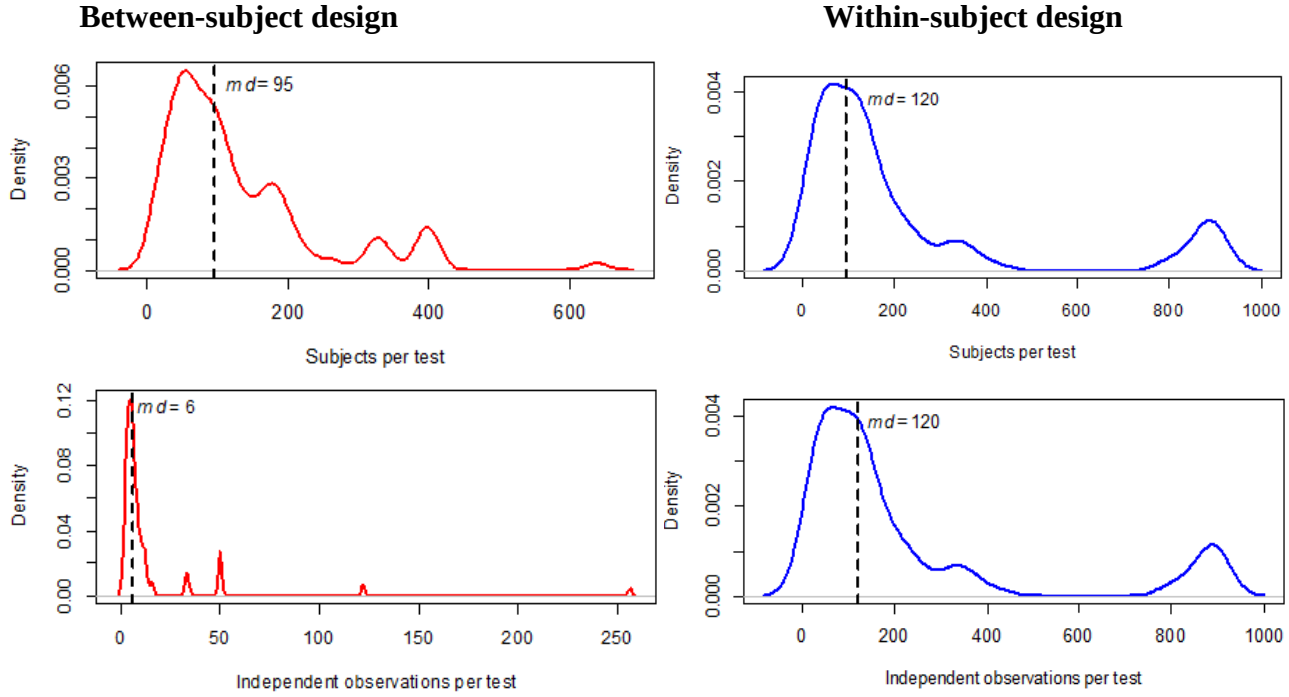

Figure 1. Subjects and independent observations per test.
Notes: The upper panel of Figure 1 shows the number of subjects per test in between-subject designs on the left side and within-subject designs on the right side. The lower panel of Figure 1 shows the number of independent observations per test in between-subject designs on the left side and within-subject designs on the right side. Source: *Experimental Economics*, *Journal of Economic Behavior and Organization* and *Games and Economic Behavior*, 2011-2017. See the list of laboratory experiments included for the analysis in the appendix.

A vast majority of the experiments (around 80%) use between-subject design randomizing between sessions, we at first analyze the power of experiments in this case. We collected data on the number of sessions for each test thus we can analyze the power of experiments given the observed range of the number of sessions. We distinguish between a small, and for experiments realistic, effect size of 0.2 (times the standard deviation) and a medium, and ambitious, effect size of 0.5 (Cohen, 1988). For each session, we use the median number of 21 subjects calculated from our sample. More importantly, we consider the intra-cluster correlation of the data between sessions for power calculations (Spybrook et al., 2011). The elementary power calculations are summarized in Table 3.

Table 3. Power calculations depending on sessions number, effect size, intra-cluster correlation

Effect size		**0.2**				**0.5**			
ICC / **N of sessions**		0	0.1	0.25	0.5	0	0.1	0.25	0.5
Minimum	2	-	-	-	-	-	-	-	-
Median	10	0.24	0.12	0.08	0.06	**0.89**	0.46	0.26	0.17
Maximum	68	**0.95**	0.58	0.33	0.21	**0.99**	**0.99**	**0.96**	**0.81**

Notes: N of sessions: Number of sessions; ICC: Inter-cluster correlation. We observe the number of sessions and provide power calculations using the Optimal Design software (Raudenbush et al., 2011), two treatments, and a significance level of 0.05.

It is barely reasonably to provide any power calculations when only two sessions are run with treatment randomization on the session level – that is the minimum number of sessions that we observe – or three sessions with one treatment per session – that we also observe - as it is unclear how to differentiate the treatment effect from the session effect in those cases. Thus, we next focus on the median experiment using the median number of sessions per test from our sample. The median number of sessions per test is ten. With an effect size of 0.2, experiments with ten sessions will not have higher power than 24%.[5] A power of more than 80% can only be achieved with an ambitious effect size of 0.5 of the standard deviation and if there would be no intra-cluster correlation.[6] Only with the maximum number of sessions in our sample of 68, 80% power to find effect sizes of 0.5 is reached with all levels of intra-cluster correlation that

[5] We aim to estimate the upper bound of the statistical power. Thus, we assume two treatments per experiment.

[6] To get a sense of intra-cluster (session) correlation (ICC) for experimental economics, we estimate ICC using data from the "Experimental Economics Replication Project" (Camerer et al., 2016), which provides high-quality replications of 18 experiments from high-ranking economic journals. Based on these estimations, we see that ICC can be close to zero, but can reach more than 0.7. As these studies are provided in line with common study protocols by highly trained teams of researchers, one can assume that the level of ICC depends more on the type of outcome variable or subject pool rather than experimental procedures in these experiments. In these cases, we are likely to observe lower-bound of ICC, and thus, the assumption of zero ICC is barely justifiable for any individual experiment.

we analyze (from 0 to 0.5). Effect sizes of 0.2 will only be detected with a power of 80% if the intra-cluster correlation is below or equal to 0.04.[7]

We further provide power estimations for within-subject designs (a minority in our sample) with randomization at the individual level. We use the range of subjects of our sample, the significance level of 0.05, and distinguish between a small, and for experiments realistic, effect size of 0.2 (times the standard deviation) and a medium, and ambitious, effect size of 0.5 (Cohen, 1988). The medium number of subjects of within-subject experiments in our sample is 120 - the power to find effect sizes of 0.2 in this sample is only 58%. Effect sizes of 0.5 are found with a power of 80% and more than 130 subjects. We see that even within-subject design the studies are rarely reach desired power, but this is a minority of the studies.

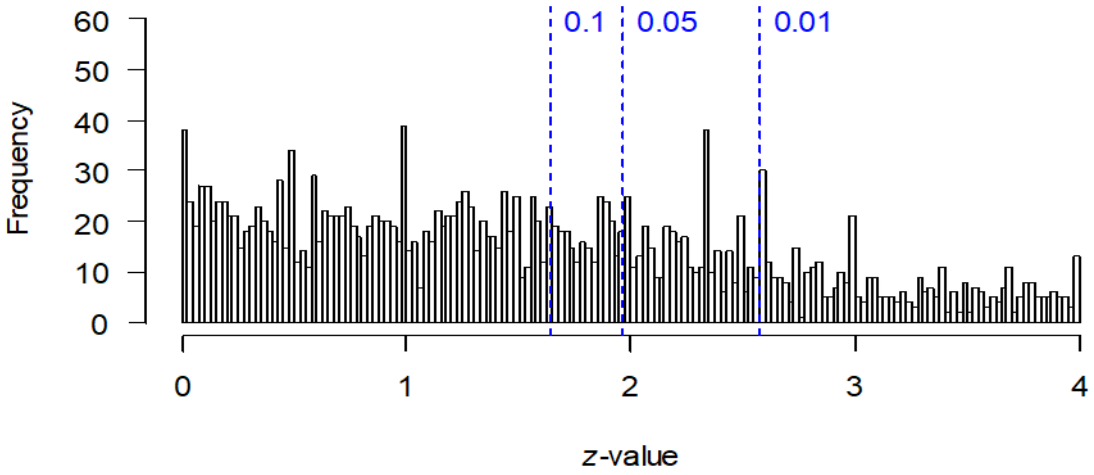

Figure 4. Histogram of *z*-values with $c = 0.025$.
Notes: The figure shows the histogram for z-values from 1 to 4. Thresholds of statistical significance are shown for $\alpha = 0.1$ ($z = 1.64$), 0.05 ($z = 1.96$) and 0.01 ($z = 2.58$).
Source: *Experimental Economics, Journal of Economic Behavior and Organization* and *Games and Economic Behavior*, 2011-2017 (see appendix).

Next, we turn to the analysis of the distribution of z-values from tests based on 3 172 tests in our sample (69 randomly selected articles) to get a sense if the analysis of laboratory experiments shows signs of the reporting bias.[8] Figure 4 compares if the papers in our sample report more p-values that are just below – vs. just above – the common thresholds of 10%, 5%,

[7] Note that more than 75% of the tests in our sample comes from parametric regression analyses. Thus, we focus on power based on parametric tests. However, a considerable minority (around 21% in our sample) is non-parametric. As non-parametric tests are more conservative, one would expect that studies can have even less power on average.
[8] We do a similar analysis as Brodeur et al. (2016) who examine the reporting bias in three journals - American Economic Review, Quarterly Journal of Economics, The Journal of Political Economy – their sample of laboratory experiments is based on 3 503 tests (86 articles).

and 1%. We analyze if the corresponding z-values in a pre-defined range above 1.64, 1.96, and 2.58 are overrepresented compared to the same range below these thresholds (caliper test). Figure 4 considers the range of the caliper size of 0.025 and shows that suspiciously too many tests in our sample are just highly significant. We investigate this suspicious pattern with the help of caliper tests – one-sided binomial tests of the z-value distribution – for different caliper sizes (c = 0.01, c = 0.025, and c = 0.05). We construct lower bound confidence levels using a non-parametric bootstrap procedure clustered on the paper level to account for the non-independence of reporting in papers (1000 bootstrap samples). In absence of a reporting bias, the number of z-values under or over the caliper (p-value = 1%, z-value = 2.58) shall be binomially distributed (p_b=0.5), especially in narrow intervals of the caliper size. However, Table 3 shows that the binomial probability exceeds 0.5 for all levels of caliper sizes, and this binomial probability is particularly large for narrow levels of caliper sizes.

Table 4. Caliper tests at the 0.01 significance level

Caliper Size c	Number of z-values under caliper	Number of z-values over caliper	Binomial probability p_b	0.95 lower confidence bound
0.010	9	30	**0.769**	**0.611**
0.025	23	46	**0.667**	**0.531**
0.050	55	68	**0.553**	0.438

Notes: Caliper tests with z-values of two-sided tests for various subsamples at the 0.01 significance level and three caliper sizes (c = 0.01, c = 0.025, and c = 0.05) are shown. Binomial probabilities p_b larger than 0.5 and 95% lower confidence bound higher than 0.5 are highlighted in bold. The confidence bound of the binomial probability is constructed with the help of a cluster-bootstrap procedure with 1000 bootstrap samples clustered at the paper level. The power of the caliper tests for the sample is 80% for the effect size of 0.25 standard deviation with ICC = 0.11. We analyze 69 papers with on average 46 tests, Intra-cluster (paper) correlation within the sample is 0.11(Lower 95% CI: 0.077; Upper 95% CI: 0.16).
Source: Experimental Economics, Journal of Economic Behavior and Organization and Games and Economic Behavior, 2011-2017 (see appendix).

We provide exploratory analysis if the reporting bias observed in Figure 4 and Table 3 cannot be explained by random, and we investigate deeper the observed strong discontinuity in our sample by the paper- and test-specific variables that we collected (see Tables 1 and 2). The estimations are reported in the appendix (Table 3A). The observed discontinuities are observed in all years. However, discontinuities especially prevalent in double-authored papers (see Beck et al., 2018, for a similar result on lying in a group of two), in Experimental Economics and Games and Economic Behavior, in main results, without eye-catchers, without hypotheses, and

in experiments with more than one round and with less than eight independent observations. Those results point out some heterogeneity in reporting bias, however, further research is needed to make a more conclusive statement. Table 4 presents a summary of our results.

Table 5. Results of hypotheses

Hypothesis	Support	Comments
Main analysis		
1a: The median laboratory experiment does not achieve the minimum desired level of statistical power.	Yes	
1b: The median laboratory experiment that randomizes at the session-level has lower statistical power than the median experiment that randomizes at the individual level.	Yes	
2a: The incidence of reporting biases is present around the 0.01 significance level.	Yes	
2b: The incidence of reporting biases is present around the 0.05 significance level.	No	
2c: The incidence of reporting biases is present around the 0.1 significance level.	No	
Exploratory analysis		
3a: The incidence of reporting biases decreases over time.	No	Evidence for reporting bias in papers published in 2013 and 2015.
3b: The higher the number of authors that are involved with the research and the article, the lower the levels of reporting biases.	No	Evidence for reporting bias in double-authored papers.
3c: The incidence of reporting biases varies with the journal.	Yes	Evidence for reporting bias in *Experimental Economics* and *Games and Economic Behavior*.
3d: The incidence of reporting biases is higher among the main results than among non-main results.	Yes	Evidence for reporting bias in the main results.
3e: The incidence of reporting biases is higher among tests denoting statistical significance merely through eye-catchers than among tests explicitly declaring p-values.	No	Strong evidence for reporting bias in tests without eye-catchers.
3f: Articles without an explicit list of hypotheses are more vulnerable to biased reporting.	Yes	Evidence for reporting bias in tests without an explicit list of hypotheses.
3g: Articles without a theoretical model are more vulnerable to biased reporting.	Yes	Evidence for reporting bias in articles without a theoretical model.
3h: The incidence of reporting biases is higher if the experiment uses repeated games than one-shot games.	Yes	Evidence for reporting bias in repeated games.

3i: The incidence of reporting biases is influenced by the number of independent observations.	No	Weak evidence for reporting bias in tests with less than 8 independent observations.

6. Conclusion and Discussion

We drew a random sample of lab experiments published in three specialty journals from 2011-2017 and analyzed their number of independent observations, their power, and indications of reporting biases. We find that the median experiment in our sample uses between-session randomization with 10 independent observations yielding to the power of typically less than 24%. Furthermore, we find evidence for reporting biases for highly significant results ($p < 0.01$), which are especially noticeable in papers without hypotheses.

Looking at our results, we are still at the beginning to find a way out of the replicability and credibility crisis in economics. In this paper, we draw attention to the unit of randomization as it determines effective sample size, independent observations, and thus, the power of the experiment. Specifically, we wanted to assess if those design effects are taken into account in the laboratory experiments at the stage of execution of the experiment. The evidence-based on not large, but a random sample of papers in three specialty journals from 2011-2017 points out that it is unlikely to be the case. The vast majority of the experiments randomize on the session level resulting in too few independent observations, thus, undermines the power. Moreover, we find that laboratory experiments also not immune to reporting biases.

Though disappointing, those results have a straightforward positive message. The researchers shall pay more attention to the level of randomization in the laboratory experiments (see also Bellemare et al., 2014, and Kim, 2020). For instance, run experiments with a large number of sessions e.g. a large number of sessions but with a small number of participants in each session. Alternatively, the experimenters can randomize within sessions, i.e. run all treatments within one session, and they consider balancing their randomization on observables (Bruhn and McKenzie, 2009) – e.g., concerning gender, age or data obtained from a pre-experimental questionnaire. If within-session randomization or balanced randomization is not an option, one can generate control variables using a double-lasso selection procedure on a pre-specified set of variables (Belloni et al., 2014), but those control variables shall be measured in a pre-experimental questionnaire. The more powerful approach is to include a pre-treatment measure of the outcome variable as a control in the specification – this approach can be implemented in the laboratory without complications (see for example Asanov and Vannuccini,

2020). Moreover, Pre-registering experiments with ex-ante power calculations, pre-analysis plans, and a set of hypotheses before starting the experiment are good tools to alleviate problems discussed in our paper and should be used more often (Christensen and Miguel, 2018; Burlig, 2018). Finally, researchers should be encouraged to share and combine their data to obtain more powerful experiments (Button et al., 2013). For instance, collaborations in the open science framework show the fruitfulness of this research paradigm (Anderson, 2015; Klein et al., 2014; Klein et al., 2018).

References

Abadie, A. (2020) 'Statistical Nonsignificance in Empirical Economics', *American Economic Review: Insights*, 2(2), 193-208.

Anderson, J. E., Aarts, A. A., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., ... and Bartmess, E. (2015) 'Estimating the reproducibility of psychological science', *Science*, 349(6251). doi: 10.1126/science.aac4716.

Andrews, I., and Kasy, M. (2019) 'Identification of and correction for publication bias', *American Economic Review*, 109(8), 2766-94.

Asanov, I. and Vannuccini, S., (2020) 'Short-and long-run effects of external interventions on trust', *Review of Behavioral Economics*, Vol. 7: No. 2015-013. http://dx.doi.org/10.1561/105.00000118

Baker, M. (2016) 'Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the 'crisis rocking science and what they think will help', *Nature*, 533, pp. 452–454. doi: 10.1038/533452a.

Bastardi, A., Uhlmann, E. L. and Ross, L. (2011) 'Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence', *Psychological Science*, 22(6), pp. 731–732. doi: 10.1177/0956797611406447.

Beck, T., Bühren, C., Frank, B., and Khachatryan, E. (2018) 'Can Honesty Oaths, Peer Interaction, or Monitoring Mitigate Lying?', *Journal of Business Ethics*, pp. 1-18. doi: 0.1007/s10551-018-4030-z.

Bellemare, C., Bissonnette, L. and Kröger, S. (2014) *Statistical Power of within and Between-Subjects Designs in Economic Experiments, IDEAS Working Paper Series from RePEc*. doi: 10.2139/ssrn.3149007.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014) "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28 (2): 29-50.

Benjamin, D., J. Berger, M. Johannesson, B. Nosek, E.-J. Wagenmakers, R. Berk, K. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. Chambers, M. Clyde, T. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. Field, M. Forster, E. George, R. Gonzalez, S. Goodman, E. Green, D. Green, A. Greenwald, J. Hadfield, L. Hedges, L. Held, T.-H. Ho, H. Hoijtink, J. Jones, D. Hr- uschka, K. Imai, G. Imbens, J. Ioannidis, M. Jeon, M. Kirchler, D. Laib- son, J. List, R. Little, A. Lupia, E. Machery, S. Maxwell, M. McCarthy, D. Moore, S. Morgan, M. Munaf, S. Nakagawa, B. Nyhan, T. Parker, L. Per- icchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. Schnbrodt, T. Sel- lke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. Watts, C. Winship, R. Wolpert, Y. Xie, C. Young, J. Zinman, and V. Johnson (2018) 'Redefine Statistical Significance', *Nature Human Behaviour*, 2, 6-10.

Bergh, D. D., Sharp, B. M., Aguinis, H., and Li, M. (2017) 'Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings', *Strategic Organization*, 15(3), pp. 423-436. doi: 10.1177/1476127017701076.

Berry, J., L. C. Coffman, D. Hanley, R. Gihleb, and A. J. Wilson (2017) 'Assessing the rate of replication in economics', *American Economic Review*, 107, 27- 31.

Biau, D. J., Jolles, B. M. and Porcher, R. (2010) 'P value and the theory of hypothesis testing: An explanation for new researchers', *Clinical Orthopaedics and Related Research*, 468(3), pp. 885–892. doi: 10.1007/s11999-009-1164-4.

Brodeur, B. A. *et al.* (2016) 'Star Wars: The Empirics Strike Back', *American Economic Journal: Applied Economics*, 8(1), pp. 1–32. doi: 10.1257/app.20150044.

Bruhn, M. and McKenzie, D. (2009) 'In Pursuit of Balance: Randomisation in Practice in Development Field Experiments', *American Economic Journal: Applied Economics*, 1(4),

pp. 200–232. doi: 10.1257/app.1.4.200.
Bruns, S. B. *et al.* (2019) 'Reporting errors and biases in published empirical findings: Evidence from innovation research', *Research Policy*, 48(9), p. 103796. doi: 10.1016/j.respol.2019.05.005.
Burlig, F. (2018). Improving transparency in observational social science research: A pre-analysis plan approach. *Economics Letters*, *168*, pp. 56-60. doi: 10.1016/j.econlet.2018.03.036.
Butera, L. and J. A. List (2017) *'An Economic Approach to Alleviate the Crisis of Confidence in Science: With an Application to the Public Goods Game', NBER Working Papers*, 23335.
Butera, L., Grossman, P. J., Houser, D., List, J. A., & Villeval, M. C. (2020) *'A New Mechanism to Alleviate the Crises of Confidence in Science-With An Application to the Public Goods Game' NBER Working Papers*, 26801.
Button, K. S. *et al.* (2013) 'Power failure: why small sample size undermines the reliability of neuroscience.', *Nature reviews. Neuroscience*. Nature Publishing Group, 14(5), pp. 365–76. doi: 10.1038/nrn3475.
Byington, E. K., and Felps, W. (2017) 'Solutions to the credibility crisis in management science', *Academy of Management Learning and Education*, 16(1), pp. 142-162. doi: 10.5465/amle.2015.0035.
Card, D. and Dellavigna, S. (2013) 'Nine Facts about Top Journals in Economics', *Journal of Economic Literature*, 51(1), pp. 144–161. doi: 10.1257/jel.51.1.144.
Camerer, C. F. *et al.* (2016) 'Evaluating replicability of laboratory experiments in economics', *Science*, 351(6280), pp. 1433–1436. doi: 10.1126/science.aaf0918.
Camerer, C., A. Dreber, F. Holzmeister, T. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, W. E.-J., and H. Wu (2018) 'Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015', *Nature Human Behaviour*, 2, 637–644. https://doi.org/10.1038/s41562-018-0399-z
Christensen, G., & Miguel, E. (2018) 'Transparency, reproducibility, and the credibility of economics research.' *Journal of Economic Literature*, *56*(3), pp. 920-80. doi: 10.1257/jel.20171350.
Coffman, L. C., M. Niederle, and A. J. Wilson (2017) 'A Proposal to Organize and Promote Replications', *American Economic Review: Papers and Proceedings*, 107, 41-45.
Cohen, J. (1988) *Statistical power analysis for the behavioral science*s. 2nd. ed. Hillsdale: Erlbaum Associates.
Czibor, E., Jimenez-Gomez, D. and List, J. A. (2019) 'The Dozen Things Experimental Economists Should Do (More Of)', *SSRN Electronic Journal*, pp. 1–74. doi: 10.2139/ssrn.3313734.
Drazen, A., A. D. Almenberg, E. Y. Ozbay, and E. Snowberg (2019) *'A Journal- Based Replication of Being Chosen to Lead', NBER Working Papers*, 26444.
DellaVigna, S, and E. Linos (2020) '*RCTs to Scale: Comprehensive Evidence from Two Nudge Units. Working Paper*, UC Berkeley.
Dreber, A., d. A. J. Pfeiffer, Thomas a, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson (2015) 'Using Prediction Markets to Estimate the Reproducibility of Scientific Research', *Proceedings of the National Academy of Sciences*, 112, 15343-15347.
Fanelli, D., Costas, R. and Ioannidis, J. P. A. (2017) 'Meta-assessment of bias in science', *Proceedings of the National Academy of Sciences*, 114(14), pp. 3714–3719. doi: 10.1073/pnas.1618569114.
Fischbacher, U. (2007) 'Z-Tree: Zurich toolbox for ready-made economic experiments', *Experimental Economics*, 10(2), pp. 171–178. doi: 10.1007/s10683-006-9159-4.
Fischbacher, U., Bendrick, K. and Schmid, S. (2015) *z-Tree 3.5 Tutorial and Reference Manual*,

Zurich Toolbox for Readymade Economic Experiments (z-Tree).
Fisher, R. A. (1925) *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
Fisher, R. A. (1932) *Statistical methods for research workers*. 5th ed. Statistical methods for research workers: Oliver and Boyd.
Franco, A., N. Malhotra, and G. Simonovits (2014) 'Social science. Publication bias in the social sciences: unlocking the file drawer Science', 345, 1502-1505.
Gerber, A. S. and Malhotra, N. (2008) 'Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals', *Quarterly Journal of Political Science*, 3(3), pp. 313–326. doi: 10.1561/100.00008024.
Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016) 'Comment on "Estimating the reproducibility of psychological science'. *Science*, 351(6277), 1037-1037. doi: 10.1126/science.aad7243.
Glennerster, R. and Takavarasha, K. (2013) *Running randomized evaluations: a practical guide*. Princeton University Press.
Goodman, S. (2008) 'A Dirty Dozen: Twelve P-Value Misconceptions', in *Seminars in Hematology*. WB Saunders, pp. 135–140. doi: 10.1053/j.seminhematol.2008.04.003.
Goodman, S. N. (1999) 'Toward evidence-based medical statistics. 1: The P value fallacy', *Annals of internal medicine*, 130(12), pp. 995–1004. doi: 10.7326/0003-4819-130-12-199906150-00008.
Goodman, S. N., Fanelli, D. and Ioannidis, J. P. A. (2016) 'What does research reproducibility mean?', *Science Translational Medicine*, 8(341), p. 341ps12. doi: 10.1126/scitranslmed.aaf5027.
Greenland, S. *et al.* (2016) 'Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations', *European Journal of Epidemiology*. Springer Netherlands, 31(4), pp. 337–350. doi: 10.1007/s10654-016-0149-3.
Haahr, M. (2019) *RANDOM.ORG: True Random Number Service*. Available at: https://www.random.org/ (Accessed: 23 July 2019).
Hubbard, R. and Lindsay, R. M. (2008) 'Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing', *Theory & Psychology*, 18(1), pp. 69–88. doi: 10.1177/0959354307086923.
Hudson, J. (1996) 'Trends in Multi-Authored Papers in Economics', *Journal of Economic Perspectives*, 10(3), pp. 153–158. doi: 10.1257/jep.10.3.153.
Ioannidis, J. P. A. (2005) 'Why Most Published Research Findings Are False', *Chance*, 18(4), pp. 40–47. doi: 10.1080/09332480.2005.10722754.
Ioannidis, J. P. A. (2008) 'Why most discovered true associations are inflated.', *Epidemiology*, 19(5), pp. 640–648. doi: 10.1097/EDE.0b013e31818131e7.
Ioannidis, J. P. A., Stanley, T. D. and Doucouliagos, H. (2017) 'The Power of Bias in Economics Research', *Economic Journal*, 127(605), pp. F236–F265. doi: 10.1111/ecoj.12461.
Ioannidis, J. P. A. and Trikalinos, T. A. (2007) 'An exploratory test for an excess of significant', *Clinical trials*, 4(3), pp. 245–253. doi: 10.1177/1740774507079441.
Kerr, N. L. (1998) 'HARKing: Hypothesizing After the Results are Known', *Personality and Social Psychology Review*, 2(3), pp. 196–217. doi: 10.1207/s15327957pspr0203_4.
Kim, D. (2020). Clustering Standard Errors at the 'session' level. mimeo. https://kimdukgyoo.github.io/PDFfiles/ClusteringSEsession.pdf
Klein, R.A., et al. 2014. Investigating variation in replicability: A ''many labs'' replication project. *Social Psychology*, 45(3), pp. 142–152. doi: 10.1027/1864-9335/a000178.
Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), pp. 443-490. doi: 10.1177/2515245918810225.

Kunda, Z. (1990) 'The Case for Motivated Reasoning', *Psychological Bulletin*, 108(3), pp. 480–498. doi: 10.1037/0033-2909.108.3.480.
Maniadis, Zacharias, Fabio Tufano, and John A. List (2017) 'To Replicate or Not To Replicate? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study', *Economic Journal* 127, no. 605.
McKenzie, D., 2012. Beyond baseline and follow-up: The case for more T in experiments. *Journal of development Economics*, *99*(2), pp.210-221.
McCloskey, D. N. and Ziliak, S. T. (1996) 'The Standard Error of Regressions', *Journal of Economic Literature*, 34(1), pp. 97–114. Available at: http://www.jstor.org/stable/2729411.
Mead, R., Gilmour, S. G. and Mead, A. (2012) *Statistical principles for the design of experiments : applications to real experiments*. Cambridge: Cambridge University Press.
Moonesinghe, R., Khoury, M. J. and Janssens, A. C. J. W. (2007) 'Most Published Research Findings Are False—But a Little Replication Goes a Long Way', *PLoS Medicine*, 4(2), p. e28. doi: 10.1371/journal.pmed.0040028.
Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2019) 'Replication studies in economics—How many and which papers are chosen for replication, and why?' *Research Policy*, 48(1), pp. 62-83. doi: 10.1016/j.respol.2018.07.019.
Munafò, M. R. *et al.* (2017) 'A manifesto for reproducible science', *Nature Human Behaviour*. Macmillan Publishers Limited, 1(1), pp. 1–9. doi: 10.1038/s41562-016-0021.
Pashler, H. and Harris, C. R. (2012) 'Is the Replicability Crisis Overblown? Three Arguments Examined', *Perspectives on Psychological Science*, 7(6), pp. 531–536. doi: 10.1177/1745691612463401.
Raudenbush, S. W. *et al.* (2011) 'Optimal Design Software for Multi-level and Longitudinal Research (Version 3.0) [Computer Software]'. Available at: http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od.
Renkewitz, F., Fuchs, H. M. and Fiedler, S. (2011) 'Is there evidence of publication biases in JDM research?', *Judgment and Decision Making*, 6(8), pp. 870–881.
Rosenthal, R. (1979) 'The file drawer problem and tolerance for null results.', *Psychological Bulletin*, 86(3), pp. 638–641. doi: 10.1037/0033-2909.86.3.638.
Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011) 'False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological Science*, 22(11), pp. 1359–1366. doi: 10.1177/0956797611417632.
Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014) 'P-curve: A key to the file-drawer', *Journal of Experimental Psychology: General*, 143(2), pp. 534–547. doi: 10.1037/a0033242.
Spybrook, J. *et al.* (2011) 'Optimal Design Plus Empirical Evidence: Documentation for the "Optimal Design" Software Version 3.0.', *William T. Grant Foundation*, pp. 1–215. doi: 10.1037/h0065543.
Vidgen, B. and Yasseri, T. (2016) 'P-values: misunderstood and misused', *Frontiers in Physics*, 4, p. Frontiers in Physics, 01 March 2016. doi: 10.3389/fphy.2016.00006.
Wasserstein, R. L. and Lazar, N. A. (2016) 'The ASA's Statement on p -Values: Context, Process, and Purpose', *The American Statistician*. Taylor & Francis, 70(2), pp. 129–133. doi: 10.1080/00031305.2016.1154108.
White, H. (2000) 'A Reality Check for Data Snooping', *Econometrica*, 68(5), pp. 1097–1126. doi: 10.1111/1468-0262.00152.
Zhang, L. and Ortmann, A. (2013) *Exploring the Meaning of Significance in Experimental Economics*, *Australian School of Business Working Paper*. 32. doi: dx.doi.org/10.2139/ssrn.2356018.
Ziliak, S. T. and McCloskey, D. N. (2004) 'Size matters: The standard error of regressions in the American Economic Review', *Journal of Socio-Economics*, 33(5), pp. 527–546. doi:

10.1016/j.socec.2004.09.024.
Ziliak, S. T. and McCloskey, D. N. (2008) *The cult of statistical significance : how the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

Appendix

Table 3A. Caliper tests for various sub-samples at the 0.01 significance level

Sub-samples	c	Number of z-values under caliper	Number of z-values over caliper	binomial probability p_b	0.95 lower confidence bound
2011	0.010	1	2	**0.667**	**0.667**
	0.025	1	3	**0.750**	**0.750**
2012	0.010	1	2	**0.667**	0
	0.025	4	7	**0.636**	0.250
2013	0.010	1	9	**0.900**	**0.750**
	0.025	5	10	**0.667**	0.364
2014	0.010	1	2	**0.667**	0
	0.025	2	6	**0.750**	0.500
2015	0.010	0	6	**1**	1
	0.025	0	7	**1**	1
2016	0.010	2	7	**0.778**	0
	0.025	6	9	**0.600**	0.125
2017	0.010	3	2	**0.400**	0
	0.025	5	4	**0.444**	0.167
Single-authored	0.010	1	0	0	0
	0.025	3	1	0.250	0
Double-authored	0.010	5	19	**0.792**	0.500
	0.025	9	27	**0.750**	**0.531**
Multiple-authored (authors ≥ 3)	0.010	3	11	**0.786**	**0.545**
	0.025	11	18	**0.621**	0.429
Experimental Economics	0.010	7	20	**0.741**	**0.556**
	0.025	18	31	**0.633**	0.463
Games and Economic Behavior	0.010	1	5	**0.833**	0
	0.025	1	9	**0.900**	**0.667**
Journal of Economic Behavior and Organization	0.010	1	5	**0.833**	0.333
	0.025	4	6	**0.600**	0.200
Main results	0.010	7	30	**0.811**	**0.636**
	0.025	19	44	**0.698**	0.565
Non-main results	0.010	2	0	0	0
	0.025	4	2	0.333	0
With eye-catchers	0.010	7	10	**0.588**	0.364
	0.025	20	25	**0.556**	0.405
Without eye-catchers	0.010	2	20	**0.909**	**0.750**
	0.025	3	21	**0.875**	**0.696**
With hypotheses	0.010	4	10	**0.714**	0.455
	0.025	12	17	**0.586**	0.333
Without hypotheses	0.010	5	20	**0.800**	**0.562**
	0.025	11	29	**0.725**	**0.531**

Sub-samples	c	Number of z-values under caliper	Number of z-values over caliper	binomial probability p_b	0.95 lower confidence bound
With formal model	0.010	6	15	**0.714**	0.333
	0.025	13	26	**0.667**	0.462
Without formal model	0.010	3	15	**0.833**	**0.636**
	0.025	10	20	**0.667**	0.458
One-shot experiments	0.010	3	4	**0.571**	0
	0.025	3	7	**0.700**	**0.400**
Repeated experiments	0.010	6	26	**0.812**	**0.630**
	0.025	20	39	**0.661**	**0.510**
Independent observations ≤ 8	0.010	3	13	**0.812**	0.500
	0.025	14	22	**0.611**	0.417
Independent observations > 8	0.010	6	15	**0.714**	0.462
	0.025	8	20	**0.714**	0.455

Notes: Caliper tests with z-values of two-sided tests for various subsamples at the 0.01 significance level and two caliper sizes (c = 0.01 and 0.025) are shown. Binomial probabilities p_b larger than 0.5 and 95% lower confidence bound higher than 0.5 are highlighted in bold. The confidence bound is constructed with the help of a cluster-bootstrapped procedure with 1000 bootstrap samples clustered on the paper level.
Source: *Experimental Economics*, *Journal of Economic Behavior and Organization* and *Games and Economic Behavior*, 2011-2017.

Papers included in the analysis

1. Ackert, L. F., Kluger, B. D. and Qi, L. (2012) 'Irrationality and beliefs in a laboratory asset market: Is it me or is it you?', Journal of Economic Behavior & Organization, 84(1), pp. 278–291. doi: 10.1016/j.jebo.2012.03.014.
2. Aguiar-Conraria, L., Magalhães, P. C. and Vanberg, C. A. (2016) 'Experimental evidence that quorum rules discourage turnout and promote election boycotts', Experimental Economics, 19(4), pp. 886–909. doi: 10.1007/s10683-015-9473-9.
3. Batista, G. *et al.* (2017) 'Do investors trade too much? A laboratory experiment', Journal of Economic Behavior and Organization, 140(August 2017), pp. 18–34. doi: 10.1016/j.jebo.2017.05.013.
4. Bolton, G. and Werner, P. (2016) 'The influence of potential on wages and effort', Experimental Economics, 19(3), pp. 535–561. doi: 10.1007/s10683-015-9453-0.
5. Bonein, A. and Denant-Boèmont, L. (2015) 'Self-control, commitment and peer pressure: a laboratory experiment', Experimental Economics, 18(4), pp. 543–568. doi: 10.1007/s10683-014-9419-7.
6. Brown, A. L. and Velez, R. A. (2016) 'The costs and benefits of symmetry in common-ownership allocation problems', Games and Economic Behavior, 96, pp. 115–131. doi: 10.1016/j.geb.2016.01.008.
7. Buckert, M., Oechssler, J. and Schwieren, C. (2017) 'Imitation under stress', Journal of Economic Behavior and Organization, 139, pp. 252–266. doi: 10.1016/j.jebo.2017.04.014.
8. Cabrales, A., Nagel, R. and Rodríguez Mora, J. V. (2012) 'It is Hobbes, not Rousseau: An experiment on voting and redistribution', Experimental Economics, 15(2), pp. 278–308. doi: 10.1007/s10683-011-9300-x.

9. Cabrera, S. *et al.* (2013) ‘Splitting leagues: Promotion and demotion in contribution-based regrouping experiments’, Experimental Economics, 16(3), pp. 426–441. doi: 10.1007/s10683-012-9346-4.
10. Cárdenas, J. C. *et al.* (2014) ‘Is it my money or not? An experiment on risk aversion and the house-money effect’, Experimental Economics, 17(1), pp. 47–60. doi: 10.1007/s10683-013-9356-x.
11. Casari, M. and Luini, L. (2012) ‘Peer punishment in teams: Expressive or instrumental choice?’, Experimental Economics, 15(2), pp. 241–259. doi: 10.1007/s10683-011-9292-6.
12. Cason, T. N., Sheremeta, R. M. and Zhang, J. (2017) ‘Asymmetric and endogenous within-group communication in competitive coordination games’, Experimental Economics, 20(4), pp. 946–972. doi: 10.1007/s10683-017-9519-2.
13. Chen, Y., Jeon, G. Y. and Kim, Y.-M. (2014) ‘A day without a search engine: an experimental study of online and offline searches’, Experimental Economics, 17(4), pp. 512–536. doi: 10.1007/s10683-013-9381-9.
14. Chuah, S. H., Hoffmann, R. and Larner, J. (2014) ‘Elicitation effects in a multi-stage bargaining experiment’, Experimental Economics, 17(2), pp. 335–345. doi: 10.1007/s10683-013-9370-z.
15. Cooper, D. J. and Lightle, J. P. (2013) ‘The gift of advice: Communication in a bilateral gift exchange game’, Experimental Economics, 16(4), pp. 443–477. doi: 10.1007/s10683-012-9347-3.
16. Croson, R. *et al.* (2015) ‘Excludability: A laboratory study on forced ranking in team production’, Journal of Economic Behavior and Organization, 114(C), pp. 13–26. doi: 10.1016/j.jebo.2015.03.005.
17. Danz, D. N., Fehr, D. and Kübler, D. (2012) ‘Information and beliefs in a repeated normal-form game’, Experimental Economics, 15(4), pp. 622–640. doi: 10.1007/s10683-012-9317-9.
18. Deck, C., Servátka, M. and Tucker, S. (2013) ‘An examination of the effect of messages on cooperation under double-blind and single-blind payoff procedures’, Experimental Economics, 16(4), pp. 597–607. doi: 10.1007/s10683-013-9353-0.
19. Denant-Boemont, L., Diecidue, E. and L’Haridon, O. (2017) ‘Patience and time consistency in collective decisions’, Experimental Economics, 20(1), pp. 181–208. doi: 10.1007/s10683-016-9481-4.
20. Erat, S. and Gneezy, U. (2016) ‘Incentives for creativity’, Experimental Economics, 19(2), pp. 269–280. doi: 10.1007/s10683-015-9440-5.
21. Faillo, M., Smerilli, A. and Sugden, R. (2017) ‘Bounded best-response and collective-optimality reasoning in coordination games’, Journal of Economic Behavior and Organization, 140, pp. 317–335. doi: 10.1016/j.jebo.2017.05.015.
22. Faravelli, M. and Stanca, L. (2012) ‘When less is more: Rationing and rent dissipation in stochastic contests’, Games and Economic Behavior, 74(1), pp. 170–183. doi: 10.1016/j.geb.2011.05.008.
23. Fehrler, S. and Kosfeld, M. (2014) ‘Pro-social missions and worker motivation: An experimental study’, Journal of Economic Behavior and Organization, 100(C), pp. 99–110. doi: 10.1016/j.jebo.2014.01.010.
24. Fenig, G. and Petersen, L. (2017) ‘Distributing scarce jobs and output: experimental evidence on the dynamic effects of rationing’, Experimental Economics, 20(3), pp. 707–735. doi: 10.1007/s10683-016-9507-y.
25. Feri, F. *et al.* (2013) ‘The pivotal mechanism revisited: Some evidence on group manipulation’, Experimental Economics, 16(1), pp. 23–51. doi: 10.1007/s10683-012-9331-y.

26. Fielding, D. and Knowles, S. (2014) ‘Can you spare some change for charity? Experimental evidence on verbal cues and loose change effects in a Dictator Game’, Experimental Economics, 18(4), pp. 718–730. doi: 10.1007/s10683-014-9424-x.
27. Gächter, S., Huang, L. and Sefton, M. (2016) ‘Combining “real effort” with induced effort costs: the ball-catching task’, Experimental Economics, 19(4), pp. 687–712. doi: 10.1007/s10683-015-9465-9.
28. Gazzale, R. S. and Khopkar, T. (2011) ‘Remain silent and ye shall suffer: Seller exploitation of reticent buyers in an experimental reputation system’, Experimental Economics, 14(2), pp. 273–285. doi: 10.1007/s10683-010-9267-z.
29. Georganas, S., Levin, D. and McGee, P. (2017) ‘Optimistic irrationality and overbidding in private value auctions’, Experimental Economics, 20(4), pp. 772–792. doi: 10.1007/s10683-017-9510-y.
30. Goeree, J. K., Offerman, T. and Sloof, R. (2013) ‘Demand reduction and preemptive bidding in multi-unit license auctions’, Experimental Economics, 16(1), pp. 52–87. doi: 10.1007/s10683-012-9338-4.
31. Goertz, J. M. M. (2012) ‘Market composition and experience in common-value auctions’, Experimental Economics, 15(1), pp. 106–127. doi: 10.1007/s10683-011-9291-7.
32. Gretschko, V. and Rajko, A. (2015) ‘Excess information acquisition in auctions’, Experimental Economics, 18(3), pp. 335–355. doi: 10.1007/s10683-014-9406-z.
33. Guillen, P. and Hakimov, R. (2017) ‘Not quite the best response: truth-telling, strategy-proof matching, and the manipulation of others’, Experimental Economics, 20(3), pp. 670–686. doi: 10.1007/s10683-016-9505-0.
34. de Haan, T. and van Veldhuizen, R. (2015) ‘Willpower depletion and framing effects’, Journal of Economic Behavior and Organization, 117(C), pp. 47–61. doi: 10.1016/j.jebo.2015.06.002.
35. Hernandez-Lagos, P., Minor, D. and Sisak, D. (2017) ‘Do people who care about others cooperate more? Experimental evidence from relative incentive pay’, Experimental Economics, 20(4), pp. 809–835. doi: 10.1007/s10683-017-9512-9.
36. Jacquemet, N. and Koessler, F. (2013) ‘Using or hiding private information? An experimental study of zero-sum repeated games with incomplete information’, Games and Economic Behavior, 78(C), pp. 103–120. doi: 10.1016/j.geb.2012.12.002.
37. Jakiela, P. (2013) ‘Equity vs. efficiency vs. self-interest: On the use of dictator games to measure distributional preferences’, Experimental Economics, 16(2), pp. 208–221. doi: 10.1007/s10683-012-9332-x.
38. Jian, L., Li, Z. and Liu, T. X. (2017) ‘Simultaneous versus sequential all-pay auctions: an experimental study’, Experimental Economics, 20(3), pp. 648–669. doi: 10.1007/s10683-016-9504-1.
39. John, K. and Thomsen, S. L. (2015) ‘School-track environment or endowment: What determines different other-regarding behavior across peer groups?’, Games and Economic Behavior, 94(C), pp. 122–141. doi: 10.1016/j.geb.2015.10.007.
40. Kessler, J. B. and Norton, M. I. (2016) ‘Tax aversion in labor supply’, Journal of Economic Behavior and Organization, 124(C), pp. 15–28. doi: 10.1016/j.jebo.2015.09.022.
41. Koppel, H. and Regner, T. (2014) ‘Corporate Social Responsibility in the work place: Experimental evidence from a gift-exchange game’, Experimental Economics, 17(3), pp. 347–370. doi: 10.1007/s10683-013-9372-x.
42. Kvaløy, O. and Luzuriaga, M. (2014) ‘Playing the trust game with other people’s money’, Experimental Economics, 17(4), pp. 615–630. doi: 10.1007/s10683-013-9386-4.
43. Kvaløy, O., Luzuriaga, M. and Olsen, T. E. (2017) ‘A trust game in loss domain’, Experimental Economics, 20(4), pp. 860–877. doi: 10.1007/s10683-017-9514-7.
44. Leibbrandt, A. (2012) ‘Are social preferences related to market performance?’, Experimental Economics, 15(4), pp. 589–603. doi: 10.1007/s10683-012-9315-y.

45. Lévy-Garboua, L. *et al.* (2012) ‘Risk aversion and framing effects’, Experimental Economics, 15(1), pp. 128–144. doi: 10.1007/s10683-011-9293-5.
46. Lien, J. W., Zheng, J. and Zhong, X. (2016) ‘Preference submission timing in school choice matching: testing fairness and efficiency in the laboratory’, Experimental Economics, 19(1), pp. 116–150. doi: 10.1007/s10683-015-9430-7.
47. Lim, W., Matros, A. and Turocy, T. L. (2014) ‘Bounded rationality and group size in Tullock contests: Experimental evidence’, Journal of Economic Behavior and Organization, 99(C), pp. 155–167. doi: 10.1016/j.jebo.2013.12.010.
48. Linardi, S. (2017) ‘Accounting for noise in the microfoundations of information aggregation’, Games and Economic Behavior, 101, pp. 334–353. doi: 10.1016/j.geb.2016.05.004.
49. Maggian, V. and Villeval, M. C. (2016) ‘Social preferences and lying aversion in children’, Experimental Economics, 19(3), pp. 663–685. doi: 10.1007/s10683-015-9459-7.
50. Mimra, W., Rasch, A. and Waibel, C. (2016) ‘Price competition and reputation in credence goods markets: Experimental evidence’, Games and Economic Behavior, 100, pp. 337–352. doi: 10.1016/j.geb.2016.09.012.
51. Munro, A. and Popov, D. (2013) ‘A portmanteau experiment on the relevance of individual decision anomalies for households’, Experimental Economics, 16(3), pp. 335–348. doi: 10.1007/s10683-012-9340-x.
52. Neri, C. (2015) ‘Eliciting beliefs in continuous-choice games: a double auction experiment’, Experimental Economics, 18(4), pp. 569–608. doi: 10.1007/s10683-014-9420-1.
53. Nikiforakis, N. and Mitchell, H. (2014) ‘Mixing the carrots with the sticks: Third party punishment and reward’, Experimental Economics, 17(1), pp. 1–23. doi: 10.1007/s10683-013-9354-z.
54. Norton, D. A. and Isaac, R. M. (2012) ‘Experts with a conflict of interest: A source of ambiguity?’, Experimental Economics, 15(2), pp. 260–277. doi: 10.1007/s10683-011-9299-z.
55. Offerman, T. and Palley, A. B. (2016) ‘Lossed in translation: an off-the-shelf method to recover probabilistic beliefs from loss-averse agents’, Experimental Economics, 19(1), pp. 1–30. doi: 10.1007/s10683-015-9429-0.
56. Olivola, C. Y. and Wang, S. W. (2016) ‘Patience auctions: the impact of time vs. money bidding on elicited discount rates’, Experimental Economics, 19(4), pp. 864–885. doi: 10.1007/s10683-015-9472-x.
57. Pearson, M. and Schipper, B. C. (2013) ‘Menstrual cycle and competitive bidding’, Games and Economic Behavior, 78(C), pp. 1–20. doi: 10.1016/j.geb.2012.10.008.
58. Rietz, T. A. *et al.* (2013) ‘Transparency, efficiency and the distribution of economic welfare in pass-through investment trust games’, Journal of Economic Behavior and Organization, 94(C), pp. 257–267. doi: 10.1016/j.jebo.2012.09.019.
59. Rodriguez-Lara, I. and Moreno-Garrido, L. (2012) ‘Self-interest and fairness: Self-serving choices of justice principles’, Experimental Economics, 15(1), pp. 158–175. doi: 10.1007/s10683-011-9295-3.
60. Rosenboim, M. and Shavit, T. (2012) ‘Whose money is it anyway? Using prepaid incentives in experimental economics to create a natural environment’, Experimental Economics, 15(1), pp. 145–157. doi: 10.1007/s10683-011-9294-4.
61. Samek, A. *et al.* (2016) ‘An experimental study of the decision process with interactive technology’, Journal of Economic Behavior and Organization, 130(C), pp. 20–32. doi: 10.1016/j.jebo.2016.06.004.
62. Savikhin Samek, A. and Sheremeta, R. M. (2014) ‘Recognizing contributors: an experiment on public goods’, Experimental Economics, 17(4), pp. 673–690. doi: 10.1007/s10683-013-9389-1.

63. Sherstyuk, K., Tarui, N. and Saijo, T. (2013) ‘Payment schemes in infinite-horizon experimental games’, Experimental Economics, 16(1), pp. 125–153. doi: 10.1007/s10683-012-9323-y.
64. Shurchkov, O. (2013) ‘Coordination and learning in dynamic global games: Experimental evidence’, Experimental Economics, 16(3), pp. 313–334. doi: 10.1007/s10683-012-9339-3.
65. Stöckl, T. (2014) ‘Price efficiency and trading behavior in limit order markets with competing insiders’, Experimental Economics, 17(2), pp. 314–334. doi: 10.1007/s10683-013-9369-5.
66. Stöckl, T., Huber, J. and Kirchler, M. (2015) ‘Multi-period experimental asset markets with distinct fundamental value regimes’, Experimental Economics, 18(2), pp. 314–334. doi: 10.1007/s10683-014-9404-1.
67. Tergiman, C. (2015) ‘Institution design and public good provision: an experimental study of the vote of confidence procedure’, Experimental Economics, 18(4), pp. 697–717. doi: 10.1007/s10683-014-9423-y.
68. Tonin, M. and Vlassopoulos, M. (2013) ‘Experimental evidence of self-image concerns as motivation for giving’, Journal of Economic Behavior and Organization, 90(C), pp. 19–27. doi: 10.1016/j.jebo.2013.03.011.
69. Zhang, Y. and Du, X. (2017) ‘Network effects on strategic interactions: A laboratory approach’, Journal of Economic Behavior and Organization, 143, pp. 133–146. doi: 10.1016/j.jebo.2017.08.017.