Australian School of Business Research Paper No. 2013 ECON 32

# Exploring the Meaning of Significance in Experimental Economics

Le Zhang
Andreas Ortmann

# Exploring the Meaning of Significance in Experimental Economics

Le Zhang · Andreas Ortmann[1]

**Abstract:**

Null Hypothesis Significance Testing has been widely used in the experimental economics literature. Typically, attention is restricted to type-I-errors. We demonstrate that not taking type-II errors into account is problematic. We also provide evidence, for one prominent area in experimental economics (dictator game experiments), that most studies are severely underpowered, suggesting that their findings are questionable. We then illustrate with several examples how poor (no) power planning can lead to questionable results.

**Keywords:** Null Hypothesis Significance Testing · Type-I-errors · Type-II errors · Significance level · Statistical power

**JEL Classification** A10 · B23 ·C12

[1] L. Zhang · A. Ortmann
School of Economics, Australian School of Business, University of New South Wales, NSW 2052, Australia
Email: le.zhang1@unsw.edu.au
a.ortmann@unsw.edu.au

We thank Ben Greiner for his comments and suggestions in the Sydney Brownbag Seminar, 2013. We also thank Ben Newell, Fabio Tufano, Gigi Foster, Glenn Harrison, Leonidas Spiliopoulos, and Zacharias Maniadis for detailed comments on earlier drafts. We thank Jeanette Deetlefs for many editorial suggestions. Remaining mistakes are ours.

## 1. Introduction (Economic significance and statistical significance)

Statistical significance tends to be the gold standard when reporting empirical (experimental) findings. We call a treatment effect statistically significant when the test statistic allows us to reject the null hypothesis that there is no difference between a baseline condition and a treatment condition. Such rejection suggests (but does not prove) that the treatment effect was not a fluke. Statistical significance does not necessarily mean that the treatment effect has economic meaning, or significance. Typically, economic significance refers to the magnitude of an effect (effect size)[2]. In fact, it might be that economic significance is meaningless even though the test statistic is highly statistically significant at conventional levels.

The comparative merits of statistical significance and economic significance are vigorously contested. One school strongly objects to the usefulness of statistical testing and focuses on the importance of economic significance (McCloskey 1983; McCloskey & Ziliak 1996a; Gigerenzer 2004; McCloskey & Ziliak 2008; Ziliak & McCloskey 2008); while another school (Hoover & Siegler 2008a, b) emphasizes the importance of statistical testing as a measure for precision of estimation. Engsted (2009) agrees with the statement that statistical significance and economic significance are not the same, but questions the claim that statistical significance is useless[3]. Spanos (2008) emphasizes that it is not useful to talk about the problem without solutions; instead, he recommends severity evaluations[4] to distinguish prior- and post-data probabilities of right decisions.

In experimental economics, statistical significance is normally asserted through Null Hypothesis Significance Testing. Typically, when testing this way, researchers restrict their attention to type-I-errors. Below we demonstrate that not taking into account type-II errors is a problematic procedure. We also provide evidence, for one prominent area in experimental

[2] Equating magnitude of an effect (effect size) with economic significance is itself problematic. For example, we might find what looks like a significant effect size for a policy intervention. Whether indeed it is, depends ultimately on the benefits and costs of the intervention.

[3] He questions McCloskey and Ziliak's claim by using examples from macroeconomics and finance. He emphasizes the difference between an economic model and an econometric model (where statistical testing plays an important role). In terms of model evaluation, he summarizes two approaches: one uses statistical methods to investigate the empirical evidence of economic theory ("LSE" approach); the other uses calibrations and simulations to look at the effect size, since the misspecification of models is inherently acknowledged. Dynamic Stochastic General Equilibrium models in macroeconomics focus on quantitative information: economic insignificant but statistically significant factors in the short-term may have big effects in the long run in Vector-Auto Regression (linear rational expectations) models. For asset pricing models, Engsted (2009) also shows that the focus shifted from statistical testing to effect-size measures.

[4] The severity evaluation computes SEV (the probability of specific discrepancy) under different hypotheses about the (population) effect size: each possible outcome is mapped into the null (or alternative hypothesis) to control the probability of Type-I error and Type-II error rate.

economics (dictator game experiments), that most studies are severely underpowered, suggesting that their findings are questionable. We illustrate with several additional examples how poor (no) power planning can lead to questionable inferences. We also argue that proper power planning is a prerequisite for the determination of economic significance.

The manuscript is structured as follows: In section 2, we explain the two types of errors in detail. In section 3, we illustrate the severe situation of under-powered studies currently existing in experimental economics by calculating the statistical power for dictator game experiments. We also highlight there a few additional bad examples as well as an exemplary one. In section 4 we conclude.

## 2. Two types of errors in Null Hypothesis Statistical Testing (NHST)

### 2.1 Type-I error and Type-II error

The type-I error ("false positive") is the probability of rejecting $H_0$ at some significance level $\alpha$ when in fact $H_0$ is true. The type-II error ("false negative") is the probability of failing to reject $H_0$ at $\beta$ when in fact it is false. 1-$\beta$ is called the power of a test – it is the probability to reject the null hypothesis correctly when in fact it is false.

In order to draw reliable conclusions, we need to minimize these two types of errors. Since there is a trade-off between type-I and type-II errors[5], we cannot minimize them simultaneously, unless we increase the sample size. Take the O.J. Simpson trial[6] for example, the presumption of innocence provided the null hypothesis while the alternative hypothesis was that he was guilty as charged. The jury had to decide at what level (of accumulation of evidence) could be allowed to mistakenly reject the null hypothesis, i.e., to falsely convict Simpson if indeed he was innocent. Given what was at stake (sentencing a man to jail who was innocent, with all the costs that would involve), the jury was bound to choose a very low $\alpha$ to implement the "beyond-reasonable-doubt" provision, i.e., the jury tried to keep the type-I error small. But decreasing the probability of the type-I error mean increasing the probability of the type-II error. So the jury's dilemma was to make sure that an innocent man would not land in jail and that a person guilty as charged would. This example demonstrates that the convention of fixing a significance level is problematic. Often such levels are very contextual.

[5] The higher significance level (the probability of type-I error $\alpha$), the lower the probability of type-II error $\beta$, thus the higher the power. For instance, if result is insignificant at 5% significance level, but significant when significance level is increased to 10%, then the probability of type-II error ($\beta$) is decreased, and the power is increased.

[6] See website: http://www.intuitor.com/statistics/T1T2Errors.html

Cohen (1988) argues that power level and significance level of 0.80 and 0.05, respectively, would strike an appropriate balance of permitting a 5% chance of committing type-I errors and 20% chance of type-II errors (type-I errors are four times more serious than type-II errors). In general, it is a good idea to report both types of errors (the p-value provides the probability of making the exact type-I errors).

### 2.2 Reporting the two types of errors

Among 56 articles reviewed in Sedlmeier and Gigerenzer (1989), only two had remarks on power, with none of these providing an estimate of the power of the test. In the 1980s, only 4.4% of 181 papers in *The American Economic Review(AER)* considered the power of the test and only 1% examined the power function (McCloskey & Ziliak 1996b). Among the 95 papers published in *Experimental Economics* between 2010 and 2012, only Requate and Waichman (2011) mention statistical power and sample size issues[7]. While Cohen (1992) argues that the neglect of power may be due to a lack of understanding of the importance of power, Spanos (2008) claims that researchers did not calculate the power in *AER* because there was no clear solution to do it[8].

That the test which researchers use is the most powerful[9] may be another reason why statistical power has been neglected in empirical research. However, it is not clear whether researchers try to find the most powerful test. Furthermore, using the most powerful test does not necessarily mean that the selected test has high power, as the statistical power also depends on factors such as effect size, sample size, etc. We address this issue in more detail below.

## 3. Optimal experimental design

### 3.1 Sample size planning and the intention of experimenters

Kruschke (2011, p. 266) uses an example to explain the importance of sample size planning and the problem of "optimal stopping rule" in NHST. Assume 8 heads come up when a coin is repeatedly flipped 26 times. The p-value of getting a result as extreme as this (8 heads[10] or less, 18 tails or more) is larger than 5% if 26 flips are fixed in advance; in contrast, the p-

[7] We will explain why this neglect of power is problematic in section 3.3.

[8] We will show that while there is some truth to it, the situation is not quite as hopeless.

[9] A test is most powerful if it has greater power than any other test at the same significance level to reject the null hypothesis given that the null hypothesis is wrong.

[10] The probability of getting exact 8 heads is 0.023 ($C_{26}^{8} * 0.5^{8} * 0.5^{18}$).

value (the probability that we need to flip at least 26 times to get 8 heads[11]) is less than 5%. This example illustrates that, conditional on our assumptions about the experimenter's intentions, different conclusions can be drawn even though the data is exactly the same. If we keep collecting more data until a statistically significant result can be found (unfortunately a widely practised method that sometimes even gets encouraged), we are more likely to reject the null hypothesis that there is no treatment effect and to unduly inflate the probability of type-I errors. The bad habit of collecting more data when the current result is insignificant (but moving in the right direction), induces a larger probability of committing type-I errors. The difference can be surprisingly large. For instance, if we flip the coin up to 20 times, and stop each time to see whether the result is significant or not, the falsely rejection rate is 17.1% rather than 5% (Kruschke, 2011 p. 273)[12]. We should calculate the sample size ex ante which is powerful enough to detect an effect size of economic significance (i.e., an effect size that reflects the implied economic benefits and costs).

The practice of designing studies of low power (i.e. studies which fail to detect an effect if the effect size is "important") is questionable. If effect size is very small (as in some priming studies), we need a large sample size to identify the effect. An unduly small sample that seems to demonstrate a statistically significant effect is probably a fluke.

### 3.2 Ex-ante experimental design of optimal significance level, power level and sample size

A large sample size is preferable to minimize both type-I errors and type-II errors[13]. However, a large sample is costly. Simple rules of thumb are summarized in List *et al.* (2011): the sample size should be equal to the ratio of standard deviation in each group[14]. If experimental costs differ, the ratio of sample size should be inversely proportional to the square root of relative costs to maximise the total sample size under budget constraint. Unfortunately, these rules of thumb do not tell us how powerful the study is.

[11] The probability of getting exact 8 heads is 0.010 ($C_{25}^{7} * 0.5^{7} * 0.5^{18}$, as head is the result of the last toss).

[12] It is different from sequential probability ratio test which evaluates the whole history of data.

[13] When we use a sample to make an inference about the population, we may commit two types of mistakes: sampling errors and non-sampling errors (systematic errors). Optimal design can avoid systematic errors, but the sampling error is inevitable. The only way to minimize sampling errors is to increase our sample size. In statistics, we usually focus on large sample properties to compare different estimators: the consistency, asymptotic normality, asymptotic efficiency, central limit theory and so on.

[14] We usually hypothesize that two groups are similar and have the same standard deviation. Then the same number of participants should be recruited for each group (treatment).

Required sample size increases with confidence level and power, and decreases with effect size. Firstly, optimal sample size design requires a balance between the two types of errors. When we start to think about the importance of a study, we should consider the seriousness (i.e., the consequences) and the probability of committing either type of errors, as illustrated above by the O.J. Simpson example. For another example (of more economic relevance), the value of each life saved is claimed to be worth 6 million US dollars by the U.S. Transportation Department (Appelbaum 2011). If we can compare the costs of installation and operation of road cameras to their benefits, we can set our own α and β level and find the optimal sample size to identify the importance of cameras. Given the number of cameras fixed, we can compute how many years' data need to be included in the evaluation. A recent report[15] of the transport for NSW claims that the net cost saving is $2.294 million ($5.055-$2.762), hence the benefit cost ratio (BCR) is 1.8. This ratio can be used as a criterion for error probability ratio (β/α). If we still use 5% as our type-I error criterion, the allowance for the probability of making type-II errors ought to be at most 2.8%.

Secondly, the required sample size (or power planning) depends on the "true" effect size (the exact alternative hypothesis). We should have prior knowledge about the (expected) effect size. Some such knowledge can be found in meta-analyses for example.

### 3.3 The severe situation of under-powered studies in experimental economics

We reviewed all papers published in *Experimental Economics* for the last three years. None of the papers stated the optimal sample size design and only one identified the statistical power as an issue. As suggested by Spanos (2008), one reason for ignorance of statistical power might be the uncertainty of the exact alternative hypothesis. However, we believe that researchers working on a particular topic probably have enough prior knowledge to proceed with reasonable assumptions about the "hypothesized" population effect size (or the effect size worth detecting). An alternative is to use information embedded in meta-analyses results[16]. Below, we use the studies in Engel's (2011; see also Zhang & Ortmann 2012, 2013) meta-analysis of dictator games to illustrate the consequence of not powering up studies properly.

[15] The link is http://www.transport.nsw.gov.au/sites/default/files/b2b/publications/annual_reports/tfnsw-annual-report-2012.pdf

[16] We are aware that meta-analysis is not a panacea. It may be afflicted by various publication biases, insufficient reporting of design and implementation details, and so on.

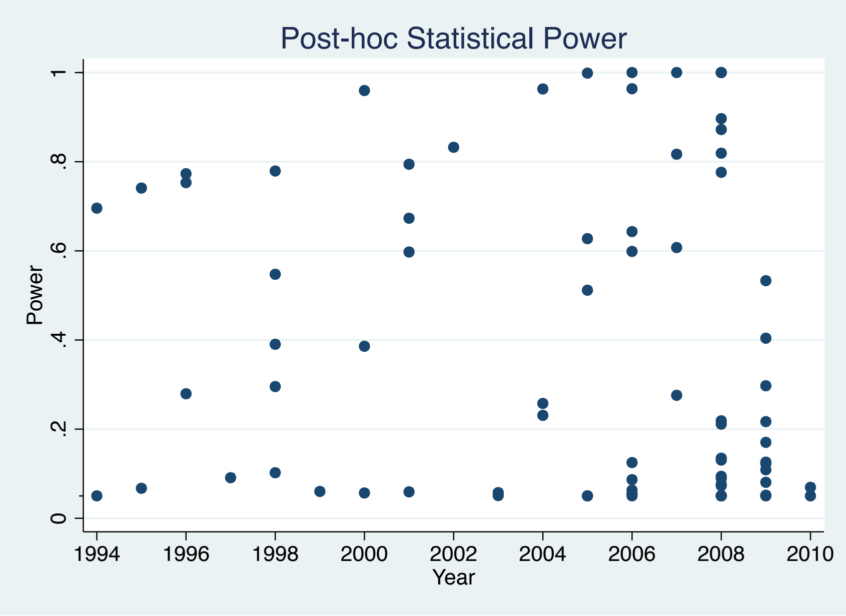

Fig 1. Post-hoc statistical power for dictator game experimental studies in Engel's (2011) meta-analysis

By using the effect sizes (marginal effects) from the meta-analysis as the true population effect size, we calculate the post-hoc statistical power for studies included in Engel's (2011) meta-analysis which investigate at least one of those explanatory variables[17] by the non-central t distribution. The lowest power is only 5% and the median power is less than 25% (average power is 38%). If the treatment effects exist, and taking the median as the relevant statistic, we only have a 25% chance to detect the effects. Fig 1 shows that the power of the studies included in Engel's meta-study varies dramatically and that there is no clear time trend pointing towards an improvement of the situation. In fact, the situation seems to have worsened over the years 2009 - 2010. This is not surprising in light of related findings from other disciplines such as neuroscience (Button *et al.* 2013), psychology (Gigerenzer *et al.* 2007); and for health-related biological and behavioural research found in Fanelli and Ioannidis (2013).

Dictator game studies test different experimental design and implementation characteristics (parameters) on giving. In order to understand whether the trend in 2009 – 2010 is driven by studies that focus on particularly susceptible design and implementation characteristics, we categorize these studies into different clusters determined by the parameters that they are studying. Most of the studies published in 2009 concern the effects of social cues, identification, and degree of uncertainty. The effect sizes for these effects are small and hence are not likely to be detected under a small sample size.

[17] This can also be done by bootstrapping samples from the data in the meta-analysis. The results are robust for either Mann-Whitney U tests or signed-rank tests.

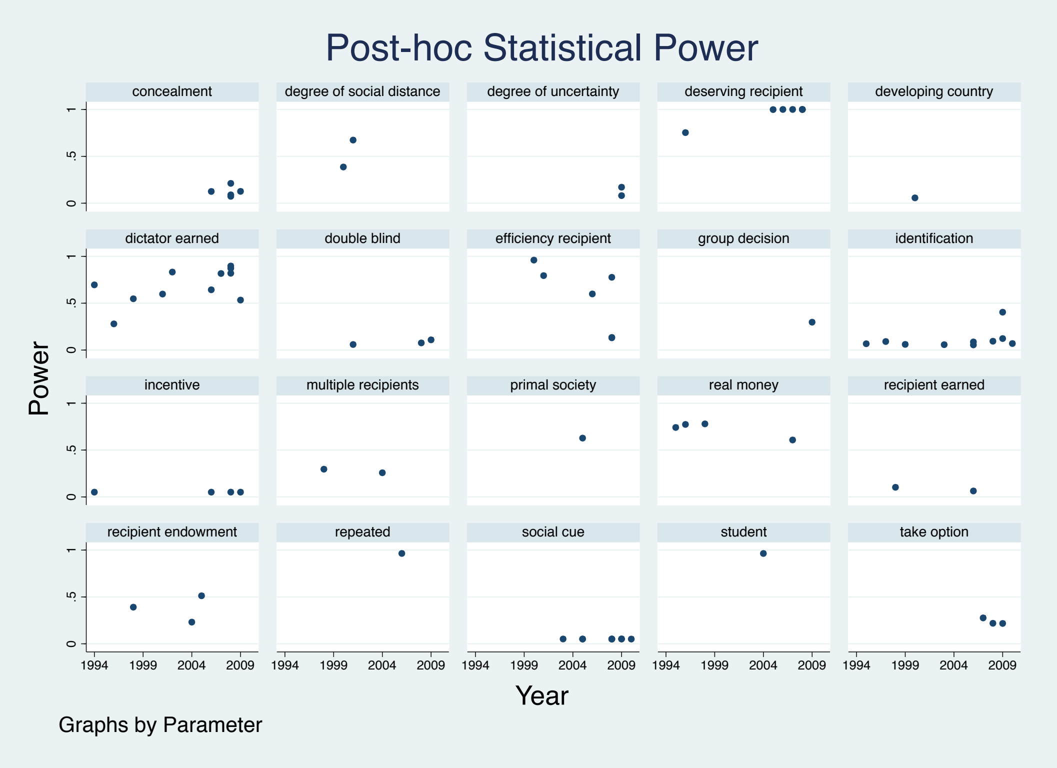

Fig 2. Post-hoc statistical power for dictator game experimental studies in Engel's (2011) meta-analysis, clustered by design and implementation characteristics

Fig 2 allows us to identify three types of studies. The first set of studies are statistically powerless (less than the median power) due to small effect sizes; the studies in this set are concerned with the effect of concealment, degree of uncertainty, double-blind, recipient earned, incentive and social cue. If the true effect size is small (despite possibly having an economic significance on charitable giving), the statistical power won't be large (smaller than the median power of 25% here), even with a reasonable sample size. The second set of studies are statistically powerful due to large effect sizes; the studies in this set are concerned with the effect of deservingness of the recipient, dictator earned (asset legitimacy) or real money. Provided that the sample size is not too small, the effects are likely to be detected. The third set comprises the remaining studies. The effect sizes for efficiency and recipient endowment are moderate, hence the statistical power of studies investigating these two factors vary dramatically due to different sample sizes.

### 3.4 Other applications of power analysis for optimal design[18]

For parametric tests, we can use Cohen's d as the effects size measure for power (sample size) planning. In experimental economics, we usually adopt non-parametric tests because the sample size is too small or the distribution is too difficult to be described by specific

[18] Here, we assume all effect sizes are worth detecting, whether big or small.

functions. Hence Cohen's d seems not to be useful (e.g., we compare medians rather than mean). As effect size is not readily available for non-parametric tests, researchers might have a ready-made excuse not to calculate power. Rosenthal and Rubin (2003) recommend that a context-free measure ($r_{\text{equivalent}}$) can be applied for non-parametric tests. In the examples below (such as Example 2), we use the $r_{\text{equivalent}}$ measure from the z-score[19] to get the effect size: $r = \frac{z}{\sqrt{N}}$ (Rosenthal & Rubin 2003), where N is the total number of observations by assuming N/2 for each groups. This r is equivalent to point-biserial correlation and can be implemented to calculate statistical power[20]. We next discuss three types of problems if statistical power is not taken into consideration.

### Question 1: how to interpret "replication failure"?

Maniadis *et al.* (2013) claim replications can dramatically improve the probability of a correct inference, even if there is only one replication. A key assumption underpinning their argument is that studies in a particular strand of research are (statistically) independent. However, replications are usually done to challenge previous evidence. For example, if the first study shows a strong effect (both economically and statistically significant), the researchers – given current editorial practices – have more incentives to work on (publish) a study that negates this finding.

There are two possible explanations for inconsistent findings (e.g., the previous literature detected treatment effects, and the treatment effects of a replication study with a smaller sample size are statistical insignificant): (1) the treatment effects are overestimated in the original study, which may be due to the notorious problem of publication bias (or a file-drawer bias) hence true effect size is likely to be overestimated); (2) replications fail to reject the null hypothesis (effect sizes are statistically insignificant) because of the low statistical power in the replication study. Indeed, Simonsohn (2013) argues that "replication failure" results from this problematic practice.

*Example 1: Market composition and experience in common-value auctions.*

[19] For example, z-scores are calculated in Mann-Whitney U test and Wilcoxon signed-rank test. In these non-parametric tests, the data is not normally distributed. However, we can compute the scores from ranks of data which are normally distributed.

[20] If population distribution is known, we can draw bootstrap samples from the population and calculate the power. This method also applies for replication studies, assuming the data from the original study reflects the true population distribution.

When not controlling for gender, Goertz (2012) fails to detect the effect of market composition in common-value auctions found in Dufwenberg *et al.* (2005). The sample size in three out of four treatments in Goertz (2012) is only half of the sample size in Dufwenberg *et al.* (2005)[21]. Her failure to detect the effect previously found may be due to the lack of statistical power[22].

*Example 2: Hidden costs of control: four repetitions and an extension.*

Ziegelmeyer *et al.* (2012) replicate the results of Falk and Kosfeld (2006) who report that hidden costs statistically significantly outweigh benefits of control (Wilcoxon signed rank tests). In the replication study, the sample size is only half of the original study - only the sample size of the extension treatment is close to the original study. It is not clear whether the different results are due to the small sample size. In Falk and Kosfeld (2006), the effect size is 0.436 (C5 treatment), -0.28(C10 treatment), 0.004(C20 treatment) by using the $r_{\text{equivalent}}$ measure. If the effect size is close to the true effect size, it is still powerful for the C5 and C10 treatments in the replication study, even though the sample size is only half of the original study[23]. However, the true effect size may be much smaller than it is in the original study; Ioannidis (2005) argues that the effect size from the first study is often exaggerated. The treatment effects are statistically more significant if we expand the dataset by a factor of two.

Maniadis *et al.* (2013) claim that the probability of a correct inference will increase with the number of replications, assuming the replication studies are (statistically) independent. If we consider the correlation between studies when undertaking consistency tests (see formula in Francis 2013), the likely overestimate of effect size and the negative relationship between the replication study and the original study aggravates the inconsistency.

**Question 2: Is the strong effect convincing? How to interpret a surprisingly strong effect?**

Previous literature does not show economic significant results consistently, but the treatment effects are both statistically and economically insignificant even with a large sample size.

[21] She implements four treatments: all-inexperienced (60 subjects), mixed-inexperienced (29 subjects), mixed-experienced (29 subjects) and all-experienced (27 subjects). The first treatment replicates the first three rounds in the previous literature, and Goertz uses the same number of observations. In the other treatments, without any explanation, the number of subjects is less than half.

[22] We could not verify our conjecture since the data are not available from the journal website.

[23] In addition to the r measure, we also draw bootstrap samples from the data of original study, and find that the results are robust (it is still powerful to detect the effect size found in the original study even though the sample size is half, given the sample is representative of the population).

From the power analysis, the effect is difficult to detect; hence, the extreme result found in the study may due to random errors, if other methods are appropriate.

*Example 3: God is watching you: priming god concepts increases prosocial behavior in an anonymous economic game.*

The meta-analysis of dictator game experiments shows that the effect size of social cues (race here) is very small. Even if we use the biggest effect size suggested by Charness and Gneezy (2008)-0.58, we need at least 30 observations for each treatment group. However, Shariff and Norenzayan (2007) recruited only 25 subjects for each treatment, which is not enough to detect a small effect at the 80% power level. It is difficult to detect the same effect (size) with the same number of observations in each group in the replication study.

**Question 3: How to interpret a treatments effects of economic significance but is statistically insignificant with a small sample size?**

If the result is statistically insignificant in the pilot paper with a small sample size, the effect size is not big yet important by some reasonable measure. We need to justify the importance of the effect size and design a new experiment with a big sample size to confirm the effect.

*Example 4: Neutral versus loaded instructions in a bribery experiment.*

Abbink and Hennig-Schmidt (2006) report that the loaded instructions do not have statistically significant effects on the average offered transfer and the frequency of permissions in a bribery experiment by using a one-sided Mann-Whitney-U test (a non-parametric test)[24]. The magnitudes of the effect sizes are small (0.042 for average transfer amount and 0.1058 for the frequency of permitting a plant), but these seemingly small effect sizes would have important real-life consequences. Assuming observed effect sizes are close to true effect sizes, we need at least 3467 and 548 observations to detect the effects with 80% power (the observed power [25] is only 0.08 and 0.15 under the current sample size). The authors will fail to detect the effects with a small sample size even if the true effects exist.

**A good example for prior power planning**

[24] All tests are one-tail test as we believe that the offer and frequency of permission is larger in neutral treatment.

[25] The observed power is different from post-hoc power, which does not assume the effect size from the sample data is the true effect size.

Previous findings for the effects of stated beliefs on experimental game play are mixed. Rutström and Wilcox (2009) argue that the mixed results may be due to low statistical power. Hence they use Monte Carlo simulations for their experimental design. They use two data-generating processes: a weighted fictitious play model with parameters estimated from Ochs's (1995) data and the 3-parameter reinforcement learning model proposed by Erev and Roth (1998). Conditional on employment 40 subject pairs for 36 periods, Rutström and Wilcox (2009) test the effects of strength of the asymmetric payoff and find that a pay-off of 19 yields a good chance of detecting the effect of stated beliefs in the Asymmetric Matching Pennies game. This example illustrates the importance of power planning and the interpretation of results.

## 4 Conclusion

We explain different concepts related to Null Hypothesis Significance Testing and the importance of appropriate statistical inference procedures in (experimental) economics. Economic significance and statistical significance answer different questions, both of which are important for the development of (social) science. An important caveat in the literature alludes to the fact that even small effects can have huge economic consequences (e.g., benefit-cost consequence), such as the example of speed cameras. It is important to make an appropriate inference from both economic significance and statistical significance. Since economic significance is easier to manipulate, it is important to evaluate the effect size (and we prefer to detect important effects). Secondly, the statistical power is as important as confidence level - the failure to detect important effects is as serious as detection of an effect which should not matter. By calculating statistical power of studies in Engel's (2011) experiments and other examples of experimental studies, we illustrated the severe situation in experimental and behavioural economics.

We provide several suggestions to researchers: firstly, pay attention to economic significance as well as statistical significance; secondly, evaluate the minimum effect size that is worth detecting and choose appropriate a, β levels and hypotheses (the null hypothesis does not have to be no treatment effect); thirdly, evaluate the estimated effect size in the literature review and use it to calculate the required (optimal) sample size if the effect is worth detecting; last but not least, report all results without discriminating statistically insignificant results (the file-drawer bias will increase inconsistency of findings). For journal editors, it is necessary not to discriminate against studies with statistically insignificant results, ceteris paribus. One way to hedge against such discrimination is the acceptance (or rejection) of a study before the data (or results) are being produced. This practice is currently being explored in other social science disciplines (Chambers & Munafo 2013).

## Appendix I: Five types of power analysis

As illustrated above, the significance level, power level, effect size and sample size are related to each other. Based on the relationship, five types of statistical power analyses (prior power analysis, post hoc power analysis, compromise power analysis, sensitivity analysis, and criterion analysis) could be done in a software called G*POWER (Faul *et al.* 2007). The prior power analysis is usually used for experimental planning, in which we calculate the required sample size with predetermined the significant level, power and population effect size. The second power analysis, post hoc power analysis, is different from the retrospective power analysis. Power is calculated basing on the sample data, significance level, sample size and population effect size, rather than use sample effect size as the population effect size to compute observed power (Faul *et al.* 2007). The compromise power analysis uses the ratio $q = \frac{\beta}{\alpha}$ and sample size to compute α and β before or after experiments. By contrast, the sensitivity is used when we want to know how much effect size would be detected under the current significance level, power level and sample size. The criterion analysis is used when the type-I error is not as important as type-II error. We use the criterion analysis to calculate α as a function of β, sample size and effect size, especially for large power and small effect size. Among all the five kinds of power analysis, the prior power analysis is the most important and fundamental analysis. It can be used for optimal experimental design.

## Appendix II: The relationship between post-study probabilities with two types of errors

The confidence level and statistical power are probabilities given that null hypothesis is true or false, but we are never know the truth. Imaging the probability of the null hypothesis being false is π (Maniadis *et al.* 2013), we can compute the post-study probability of a proper decision conditional on rejecting the null hypothesis by Bayes' rule. Ioannidis (2005) introduces two post-study probabilities: PPV, the positive predictive value, is the proportion of the real effect (null hypothesis is false) conditional on the null hypothesis being rejected in the post study; NPV, the probability of a correct conclusion (the null hypothesis is false) conditional on the failure to reject the null hypothesis.

| | $H_0$: Null is true 1-π | $H_1$: Null is false (effect exists) π | Total |
|---|---|---|---|
| Significant result (Rejection) | α(1-π) | (1-β)π | α(1-π)+(1-β)π |
| Insignificant result (Fail to reject) | (1-α)(1-π) | βπ | (1-α)(1-π)+ βπ |
| Total | (1-π) | π | 1 |

$$PPV = \frac{(1-\beta)\pi}{\alpha(1-\pi)+(1-\beta)\pi} \ and \ NPV = \frac{(1-\alpha)(1-\pi)}{(1-\alpha)(1-\pi)+\beta\pi}$$

From the table and two equations above, we can find that probability of the proper conclusion decreases by type-I errors and type-II errors. Maniadis *et al.* (2013) show that the two probabilities will be increased if there are k independent researchers competing with each other (k replications).

## References:

Abbink, K., Hennig-Schmidt, H., 2006. Neutral versus loaded instructions in a bribery experiment. Experimental Economics 9, 103-121

Appelbaum, B., 2011. As U.S. Agencies Put More Value on a Life, Businesses Fret. URL http://www.nytimes.com/2011/02/17/business/economy/17regulation.html?_r=2&

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews. Neuroscience 14, 365-76

Chambers, C., Munafo, M., 2013. Trust in science would be improved by study pre-registration

Charness, G., Gneezy, U., 2008. What's in a name? Anonymity and social distance in dictator and ultimatum games. Journal of Economic Behavior & Organization 68, 29-35

Cohen, J., 1992. A power primer. Psychological Bulletin, 155-159

Dufwenberg, M., Lindqvist, T., Moore, E., 2005. Bubbles and Experience: An Experiment. American Economic Review 95, 1731-1737

Engsted, T., 2009. Statistical vs. economic significance in economics and econometrics: further comments on McCloskey and Ziliak. Journal of Economic Methodology 16, 393-408

Falk, A., Kosfeld, M., 2006. The Hidden Costs of Control. The American Economic Review 96, 1611-1630

Fanelli, D., Ioannidis, J.P.A., 2013. US studies may overestimate effect sizes in softer research. Proceedings of the National Academy of Sciences

Faul, F., Erdfelder, E., Albert-Georg, L., Buchner, A., 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 39, 175-91

Gigerenzer, G., 2004. Mindless statistics. Journal of Socio-Economics 33, 587-606

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L.M., Woloshin, S., 2007. Helping Doctors and Patients Make Sense of Health Statistics. Psychological Science in the Public Interest (Wiley-Blackwell) 8, 53-96

Goertz, J.M.M., 2012. Market composition and experience in common-value auctions. Experimental Economics 15, 106-127

Hoover, K.D., Siegler, M.V., 2008a. The rhetoric of 'Signifying nothing': a rejoinder to Ziliak and McCloskey. Journal of Economic Methodology 15, 57-68

Hoover, K.D., Siegler, M.V., 2008b. Sound and fury: McCloskey and significance testing in economics. Journal of Economic Methodology 15, 1-37

Ioannidis, J.P.A., 2005. Why Most Published Research Findings Are False. PLoS Medicine 2, e124

List, J., Sadoff, S., Wagner, M., 2011. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. Experimental Economics 14, 439-457

Maniadis, Z., Tufano, F., List, J.A., 2013. One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. Centre for Decision Research and Experimental Economics, Discussion Paper No. 2013-07

McCloskey, D.N., 1983. The Rhetoric of Economics. Journal of Economic Literature 21, 481

McCloskey, D.N., Ziliak, S.T., 1996a. The standard error of regressions. Journal of Economic Literature 34, 97

McCloskey, D.N., Ziliak, S.T., 1996b. The Standard Error of Regressions. Journal of Economic Literature 34, 97-114

McCloskey, D.N., Ziliak, S.T., 2008. Signifying nothing: reply to Hoover and Siegler. Journal of Economic Methodology 15, 39-55

Requate, T., Waichman, I., 2011. "A profit table or a profit calculator?" A note on the design of Cournot oligopoly experiments. Experimental Economics 14, 36-46

Rosenthal, R., Rubin, D.B., 2003. r equivalent: A Simple Effect Size Indicator. Psychological Methods 8, 492-496

Rutström, E.E., Wilcox, N.T., 2009. Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. Games and Economic Behavior 67, 616-632

Sedlmeier, P., Gigerenzer, G., 1989. Do Studies of Statistical Power Have an Effect on the Power of Studies? Psychological Bulletin 105, 309-316

Shariff, A.F., Norenzayan, A., 2007. God Is Watching You: Priming God Concepts Increases Prosocial Behavior in an Anonymous Economic Game. Psychological Science (Wiley-Blackwell) 18, 803-809

Simonsohn, U., 2013. Evaluating Replication Results. SSRN, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2259879

Spanos, A., 2008. Review of Stephen T. Ziliak and Deirdre N. McCloskey's The cult of statistical significance: how the standard error costs us jobs, justice, and lives. Ann Arbor (MI): The University of Michigan Press, 2008, xxiii+322 pp. . Erasmus Journal for Philosophy and Economics 1, 154-164

Ziegelmeyer, A., Schmelz, K., Ploner, M., 2012. Hidden costs of control: four repetitions and an extension. Experimental Economics 15, 323-340

Ziliak, S.T., McCloskey, D.N., 2008. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. The University of Michigan Press, Ann Arbor.