# Evaluating Treatment Effects and Replicability*

Víctor González-Jiménez[†a] and Karl Schlag[b]

[a]Erasmus School of Economics, Erasmus University Rotterdam.

[b]Department of Economics, University of Vienna.

January 30, 2022

**Abstract**

Replication studies are important for verifying the generality of experimental investigations. In this paper we postulate incorrect testing as a novel and unexplored explanation for the lack of replicability in studies. We find that 50% of papers that failed to replicate in Camerer et al. (2016) did not have a significant treatment effect in the first place. Our evidence is based on a non-parametric test that is able to uncover treatment effects without making additional assumptions. The proposed test is powerful as it uncovers a significant effect in roughly 73% of the cases where a conventional test also detects one.

**JEL Classification:** C90, C12, C18, C14, C81

**Keywords:** experiments, treatment effects, replication, hypothesis testing, stochastic inequality test.

*We are grateful to Uri Gneezy, Roberto Weber, and participants of the NYU CESS seminar and the TIBER conference for helpful comments and discussions. We also thank the following scholars (and by extension their coauthors) who answered our questions and had willigness to share their data files: Keith Marzili Ericson, Eric Fuster, Martin Dufwenberg, Gary Charness, and Wu Hang.

[†]Corresponding author. victor.gonzalez@univie.ac.at

# 1 Introduction

Laboratory experiments generate empirical evidence for the testing of theories and identifying causal effects in different fields of economics such as mechanism design (Kagel and Levin, 2011; Dechenaux et al., 2015), labor economics (Charness and Kuhn, 2011), market design (Chen et al., 2021), and macroeconomics (Duffy, 2014). However, with the growing importance of laboratory research comes an increased care and concern for the validity of their findings.

A way to assess the validity of an empirical finding is direct replication (Nosek and Lakens, 2014). Accordingly, the conditions for obtaining a finding in a previous study are recreated to establish whether it continues to appear.[1] A prominent paper that contains 18 replication studies is Camerer et al. (2016). The findings therein have raised doubts as to the degree to which we can trust experimental findings that have not been replicated.

Our claim in a nutshell is simple. It does not make sense to check whether a finding can be replicated if the original finding itself has not been derived on a solid basis. We expand to qualify the formal grounds for this statement. Our objection is purely based on statistical hypothesis testing, the methodology of making inference in laboratory experiments. We point out that the tests used in the original studies included in Camerer et al. (2016) are not appropriate. Hence, we make a call to first uncover findings with the right methods before seeking to see whether the findings replicate in other experiments.

At the heart of the problem is the broad usage of the t-test (Student, 1908) and the incorrect interpretation of the Wilcoxon-Mann-Whitey test (Wilcoxon, 1945; Mann and Whitney, 1947) (WMW test henceforth). The former test is used to uncover whether there is a mean difference between treatments, relying however on the assumption that the underlying distribution of the random variables is approximately normal or that the underlying sample sizes tend to infinity. However, this test is not valid for comparing means of random variables for any given sample sizes if these can have any distribution (see Bahadur and Savage (1956) and Lehmann and Loh (2000)).[2]

In turn, the WMW test does not rely on the assumption of normality and is in fact non-parametric. This test is useful to uncover whether there is a difference between treatments as it tests the null hypothesis that two distributions are identical. However, it cannot be used to establish direction unless stringent assumptions on the underlying data generating process are assumed. Showing that a treatment improves or worsens the outcome of interest is the typical objective in experimental designs. This direction can be referring to, say, a treatment

[1] This type of replication is also called scientific replication (Hamermesh, 2007). We stick to the terminology used by methodologists in psychology (OSC, 2015).

[2] The reason for the discrepancy between these two statements is that there is no uniform bound. This leads to a dilemma. How large the sample has to be depends on the specific underlying data generating process.

that improved cooperation, raised effort, or increased market efficiency as compared to the control.

The wide usage of these tests to uncover treatment differences implies that experimental findings can be inadequately identified and that sample sizes can be inappropriately chosen. Replicating an experiment and using the same test to make inference does not shed more light on whether this finding is valid. To fix this problem, we use the stochastic inequality test of Schlag (2008) that is non-parametric, valid, and is able to uncover the direction of treatment effects. We call a test valid (refer to as exact in the literature) if it guarantees the type I error probability for the given sample sizes.

We apply this test to the studies featured in Camerer et al. (2016). We observe that the stochastic inequality test is powerful as it uncovers significant treatment effects for a large proportion of studies therein. Specifically, of the 22 original and replication studies that exhibit a treatment effect with a conventional test (t-test or MWM test), 16 (71%) continue to exhibit a treatment effect with the stochastic inequality test. This finding emerges despite the larger sample size requirements of the test as compared to more traditional testing alternatives which rely on distributional assumptions and/or test more stringent hypotheses.

The stochastic inequality test uncovers significant treatment effects in eight of the original studies. To assess the validity of these treatment effects, we investigate the corresponding replication data sets. We distinguish between the significance of a treatment effect in the replication data set and the difference of the magnitudes between replication and original data sets. This disctintion enables us to consider two types of replication assessments. In the first one we focus on significance. We speak of replication of the treatment effect in the *strong sense* if there is significant evidence of a treatment effect in both the original and replication study. This is similar to the standard apporach followed by Camerer et al. (2016) and OSC (2015), with the difference that we use a non-parametric valid test. Replication in the strong sense happened in four out of the eight studies that have a significant treatment effect in the original study (50%). For these studies we can confirm the authors' claims.

We note that a lack of significant evidence of the treatment effect in the replication study cannot invalidate significant findings in the original study. On the one hand, it could be that the effect is indeed not significant. On the other hand, the less stringent nature of the tested null hypothesis and the lack of distributional assumptions of the stochastic inequality test can weaken its discriminatory power. Therefore, we speak of replication in the *weak sense* if there is a significant treatment effect in the original data set and an insignificant effect in the replication data set. This happens in the data in four studies, and when the roles of original and replication studies is reversed, this number increases to seven. Since we do not find any instance in which a replication study exhibits a significant treatment effect in the opposite direction as compared

to the original study, we conclude that none of the original findings are invalidated.

Our second replication assessment also focuses on the magnitudes of the treatment effects in original and replication data sets. To that end, we use the confidence interval of the stochastic inequality test. If the confidence intervals of original and replication data sets do not intersect, our conclusion is that the treatment effect is substantially different in the replication. In these cases we say that the magnitude of the treatment effect cannot be replicated. Note that this is not necessarily bad news, if an original treatment effect is significant and the confidence interval found in the replication data set is strictly larger than that found in the original data set, the original finding is reinforced in light of new evidence. We find that the prediction does not replicate in four out of the 11 studies in Camerer et al. (2016) for which there is a significant treatment effect in original or replication data set. Among these studies one significant finding is reinforced in the replication data set, one is weakened, and there are two studies without a significant treatment effect in the replication data set. For these two insignificant studies we can state that the treatment effect exhibits considerable heteregeneity across data sets.

How do our findings compare to those of Camerer et al. (2016)? Among the studies that replicate according to their standards, we find that four (40%) strongly replicate. The claims of those studies are fully corroborated with our methodology. Furthermore, we challenge the assessment of Camerer et al. (2016) that four studies replicate. For of these studies we do not find a significant treatment effect neither in the replication nor in the original data. Here, the conclusions of both original and replication papers heavily rely on the distributional assumptions of the t-test. For the other two studies we find weak replication, i.e. the stochastic inequality test uncovers a significant treatment effect in the original but not in the replication data set. So more data would be needed to corroborate the validity of the original finding with an adequate test.

Among the six studies that do not replicate in Camerer et al. (2016), the stochastic inequality test does not uncover a significant treatment effect in the original data set in three cases. Our interpretation of this finding is that their inability to replicate is due to the lack of a significant treatment effect in the original study. Here, our methodology opens an unexplored explanation for the lack of replicability of findings, namely incorrect testing. Furthermore, for two studies that do not replicate in Camerer et al. (2016), we find that their magnitudes do not replicate, i.e. the magnitude of the treatment effect is substantially different in original and replication data sets. Hence, a reason for the lack of replicability of these studies is that the treatment effects of original and replication data sets are acutely different. It is as if they were generated by different experimental designs. This interpretation is reinforced by the comment of one of the studies in Chen et al. (2021).

# 2 Finding a Treatment Effect

## 2.1 A Brief Introduction

Laboratory experiments are a useful means to uncover causal relationships. In the simplest and typical case there are two different treatments and one wishes to investigate if there is any difference between the two of them. The MWM test enables to establish evidence of a difference, up to a prespecified significance level $\alpha$. However the WMW test is not able to identify how the difference between the two treatments is manifested. That is because that test is designed to test the null hypothesis $H_0 : F_X \equiv F_Y$ where $X$ and $Y$ are the random variables underlying the two treatments and $F_Z$ is the distribution of random variable $Z \in \{X, Y\}$. Empirical evidence to reject the null could be due to the fact that means, variances, medians, or that some other moment of their distributions are different.

Laboratory experiments are rarely designed to reveal that the distributions of two treatment arms are not identical. In the large majority of the cases the authors are interested in understanding the direction of the treatment effect. The direction is most commonly measured by comparing means or medians. To be able to compare means one needs a test of the null hypothesis $H_0 : \mathbb{E}(X) = \mathbb{E}(Y)$. To compare medians one needs to test $H_0 : med(X) = med(Y)$.

The WMW test has been used as if it were designed to test whether the means or medians of two distributions are the same. This usage of the WMW test can be justified through a *location shift model*. According to this model one assumes that $X = Y + z$ for some $z \in \mathbb{R}$ and then tests whether $z = 0$. This is equivalent to assuming $X = Y + z$ and then testing $H_0 : z = 0$ against $H_1 : z \neq 0$. However, $z \neq 0$ cannot be inferred when no assumptions are made on the outcomes apart from that they are contained in a given common bounded set.

We explain. Consider two random variables $X$ and $Y$ that generate outcomes in $[a, b]$, so $X, Y \in [a, b]$. We do not want to make assumptions on the distributions of $X$ and $Y$. Thus, regardless of which data has been gathered it could be that $P(X = a) > 0$ and $P(X = b) > 0$. As $Y = X + z$, it follows from $P(X = a) > 0$ that $z \geq 0$ as otherwise $Y < a$ would be possible. Similarly, it follows from $P(X = b) > 0$ that $z \leq 0$. Consequently we obtain that $z = 0$. So we can only infer that $z < 0$ or $z > 0$ if we can rule out that $P(X = a) * P(X = b) > 0$. However this cannot be inferred from the data set. In fact, one does not even expect to observe either $a$ or $b$ when$P(X \in \{a, b\})$ is very small. Consequently, the location shift model is useless if the random variables of interest have common bounds but are otherwise unrestricted.

In order to "rescue" the location shift model one would have to assume that $P(X \in [b - z, b]) = 0$ when $z > 0$ and $P(X \in [a, a + z]) = 0$ when $z < 0$. However, it does not seem sensible to try to find out an effect in a model that is so restrictive. The beauty of the laboratory design is that one needs to make no additional assumptions to make a true discovery about a treatment

effect, as long as the design itself ensures independence of the data.

Another widespread test used to establish treatment effects in experiments is the t-test (Student, 1908). Given the unbounded support of the normal distribution, one cannot assume that data is normally distributed when treatment outcomes are bounded. The data would be approximately normally distributed if the sample is sufficiently large for the given underlying data generating process. However, there is no uniform bound on how many observations one needs for this statement to be true (See footnote 2). Moreover, in the large majority of laboratory experiments data sets are small, hence it is hardly credible to assume that the sample is sufficiently large for the central limit theorem to kick in. Consequently, the t-test is not applicable in most laboratory experiments.

There is a valid test for identifying treatment effects in terms of means due to Schlag (2008). However this test requires substantially more data than the WMW test and the test proposed in the next section, because mean testing is very sensitive to outliers. In the next section we present an efficient alternative that builds on a different form of null hypothesis and also allows to uncover how treatment differences substantiate.

## 2.2 Testing a Stochastic Inequality and Stochastic Tendency

We present a methodology for comparing two random variables based on two independent random samples. Denote the two random variables by $X$ and $Y$. We say that $Y$ *tends to be larger* than $X$ if $P(Y > X) \geq P(Y < X)$, which holds if and only if $P(Y \geq X) \geq P(Y \leq X)$. Hence, $Y$ tends to be larger than $X$ if, whenever repeatedly drawing one value from each random variable, $Y$ realizes a higher value than $X$ more often than it realizes a lower value.

We quantify how much larger $Y$ tends to be than $X$ as follows. Specifically we introduce the stochastic tendency between X and Y as the variable $d = d(X, Y)$ where $d$ satisfies:

$$P(Y > X + d) = P(Y < X + d). \tag{1}$$

If there are multiple solutions to (1) then we define the stochastic tendency as the midpoint of them. We prove below that there is always at least one solution to (1) and that any such solution is a median of $Y - X$.

We first show that $d$ satisfies (1) if and only if $d$ is a median of $Y - X$. Let $d$ satisfy (1). From the definition of stochastic tendency we obtain $P(Y - X > d) + \frac{1}{2}P(Y - X = d) = P(Y - X < d) + \frac{1}{2}P(Y - X = d) = \frac{1}{2}$. Consequently, $P(Y - X \geq d) \geq \frac{1}{2}$ and $P(Y - X \leq d) \geq \frac{1}{2}$. So $d$ is a median of $Y - X$. Conversely, let $d'$ be some median of $Y - X$. Then $d'$ obviously satisfies (1) when $P(Y - X = d') = 0$ as in that case $P(Y - X \geq d') = 1 - P(Y - X \leq d')$ and hence $P(Y - X \geq d') = \frac{1}{2} = P(Y - X \leq d')$. Assume $P(Y - X = d') > 0$ and that $d''$ is a median such that $d'' > d'$. Then

$\frac{1}{2} \leq P(Y-X \geq d'') \leq P(Y-X > d') = 1 - P(Y-X \leq d') \leq \frac{1}{2}$. So $P(Y-X \geq d'') = P(Y-X \leq d') = \frac{1}{2}$. Hence, $P\big(Y-X \in (d', d'')\big) = 0$ and consequently any $d \in [d', d'']$ satisfies (1). This completes the proof. We now show that a solution to (1) always exists. However, this follows from the (well known) existence of medians. We recall the proof. Let $d^- = sup\Big\{d : P(Y-X \leq d) \leq \frac{1}{2}\Big\}$ and $d^+ = inf\Big\{d : P(Y-X \geq d) \leq \frac{1}{2}\Big\}$. Then $d^- = d^+$ is a median of $Y-X$.

The stochastic tendency has the same unit of measure as $X$ and $Y$. This is not true for the stochastic difference, $P(Y > X) - P(Y < X)$ (Agresti, 1984; Cliff, 1993).

We now proceed to make inference on the stochastic tendency between $X$ and $Y$. We do this based on two independent samples, $x_1, .., x_n$ and $y_1, .., y_m$, drawn from $X$ and $Y$, respectively. We first present an estimate of the stochastic tendency $d$. Let $sgn$ be the signum function, so $sgn(w) = 1$ if $w > 0$, $sgn(w) = 0$ if $w = 0$ and $sgn(w) = -1$ if $w < 0$. Define the following function

$$f(z) = \sum_{i=1}^{n} \sum_{j=1}^{m} sgn\big(y_i - x_j - z\big).$$

An estimate $\hat{d} = \hat{d}\Big(\big(x_j\big)_{j=1}^{n}, \big(y_i\big)_{i=1}^{m},\Big)$ of the stochastic tendency $d$ is given by

$$\hat{d} = \frac{1}{2} \sup\big\{z : f(z) \leq 0\big\} + \frac{1}{2} \inf\big\{z : f(z) \geq 0\big\}.$$

In words, we first produce all pairings of the two samples. Then we search for all values of $z$ such that at least half of the pairings have a value of $Y-X$ above $z$ and for at least half it lies below $z$. This yields an interval of such values. The estimate $\hat{d}$ is then given by the midpoint of this interval.

We now present one- and two-sided tests as well as confidence intervals for the stochastic tendency. We start by presenting a one-sided test. For any given $d_0$ we wish to test $H_0 : P(Y > X + d_0) \geq P(Y < X + d_0)$, alternatively this hypothesis can be written as $H_0 : d \leq d_0$. This is a test of stochastic inequality as defined by Brunner and Munzel (2000) and first implemented with a valid (finite sample) test by Schlag (2008).

We speak of a *valid test* if for the given sample sizes $n$ and $m$ it can be mathematically proven that the null hypothesis is falsely rejected with a probability that is at most $\alpha$. The statistics literature uses the term "exact" instead. We however opt for the term valid, as exact seems to infer that it is a special mathematical property. Importantly, validity formally requires a proof. On the other hand, a single counter example suffices to show that a test is not valid. Simulations cannot be used to show that a test is valid. The reason is that not all distributions can be simulated as there will be infinitely many in the null hypothesis. Hence there could be some distribution that leads to a rejection probability above alpha is overlooked.

The properties of the test suggested by Brunner and Munzel (2000) is proven using asymp-

totic theory, so assuming that $n$ and $m$ are infinitely large. As discussed above, this is an undesirable assumption for laboratory experiments. We present the valid test of $H_0 : P(Y > X) \geq P(Y < X)$ due to Schlag (2008). This can be used as a test of $H_0 : d \leq d_0$ by replacing $X$ with $X + d_0$. Let $\alpha$ be the desired significance level of the test. Let $\theta$ be a parameter. Consider first $\theta$ as given, we mention later how it is chosen. First, randomly match observations from $Y$ and $X$. Subsequently, determine whether there are significantly less pairs in which the observation from $Y$ is strictly larger than the observation from $X$ than pairs in which it is strictly smaller. This inference is done using the binomial test with level $\theta\alpha$. This matching of the entire sample into pairs is performed a large number of times after which one records the frequency at which the null was rejected. If this frequency is larger than $\theta$, then the original null hypothesis $P(Y > X) \geq P(Y < X)$ is rejected. This generates a test with level $\alpha$. The value of $\theta$ is then chosen to yield the largest set of distributions under which the rejection probability is above $\frac{1}{2}$ using the given sample sizes $n$ and $m$.[3] This test is implemented in R and STATA.[4]

The test of $H_0 : P(Y > X) \leq P(Y < X)$ is constructed analogously, simply by recording how often there are significantly less pairs in which the observations from $X$ are strictly larger than those from $Y$. Replacing $X$ with $X + d_0$ we then obtain a valid test of $H_0 : d \geq d_0$. The two-sided test of $P(Y > X + d_0) = P(Y < X + d_0)$ is constructed by combining the two one-sided tests each using $\alpha/2$ instead of $\alpha$. The null is rejected if either of them rejects the one-sided null hypothesis at level $\alpha/2$. Equi-tailed confidence intervals for the stochastic tendency with a confidence level of $1 - \alpha$ are computed by collecting all values $d_0$ under which the two-sided test cannot be rejected at level $\alpha$.

The test of stochastic inequality along with its extension to a test of stochastic tendency is powerful. It limits attention to an ordinal comparison between the two variables and hence is less sensitive to unobserved outcomes. Its power is manifested in its implementation as the test is based on the binomial test. Being an ordinal test, inference is invariant to a monotone transformation of the outcomes. Magnitudes of differences are measured in the scale of the random variables and hence are easy to interpret. In order to apply this test, one has to be willing to steer away from mean comparison, to talk about differences in tendency as opposed to differences in means

[3] For details see Schlag (2008).

[4] The R program is part of the package npExact program and can be found at https://CRAN.R-project.org/package=npExact. The STATA program can be found at https://homepage.univie.ac.at/karl.schlag/research/statistics/npStochin$_0$905.*do*. Examples of their implementation can be found at https://homepage.univie.ac.at/karl.schlag/research/statistics.

## 2.3 Replication Methodology

Replication refers to the practice of finding out whether the results obtained in an experiment continue to hold if it Is run again while preserving the conditions believed sufficient for obtaining the original finding (OSC, 2015; Nosek and Lakens, 2014). Traditionally, replicability is claimed when the replication data shows a significant treatment effect. On the other hand, lack of replicability is concluded when no significant effect is found. This is the main approach followed by OSC (2015) and Camerer et al. (2016).[5]

Lack of replicability can occur out of many reasons. It could be that the original findings are not representative or that they are less pronounced than originally claimed and emerged as a false positive. Moreover, it could also be that the replication design is not representative or that insignificance emerged as a false negative.

We highlight that evidence of lack of replicability might have nothing to do with the validity of the original effect if the test used for inference is not valid. Applying a non-valid test to the original data set might reveal a treatment effect when there is really none and hence there is nothing to replicate. This could be because the distributional assumptions of the non-valid test exaggerate the treatment effect, or because the null hypothesis of the non-valid test is misinterpreted. For similar reasons, applying a non-valid test to the replication data set might fail lead the researcher to the conclusion that a study replicates when in fact there is no treatment effect in the replication study.

Valid tests typically generate less rejections of the null than their non-valid counterparts. Consequently, valid tests require larger data sets to make useful inference. The reason why the non-valid counterparts get by with smaller sample sizes is because they either invoke additional assumptions or test more specific null hypotheses. For instance, the t-test additionally assumes that samples are normally distributed with identical variances. In turn, the WMW test investigates identity of distributions instead of the more general hypothesis that the stochastic tendency is equal to 0.

What does this lower discriminatory power of the test imply for the evidence of lack of replicability under the non-valid tests? It is unlikely that the use of a valid test reverses the findings under a non-valid test. The valid test will rarely turn the insignificant result in the replication data into a significant finding. On the contrary, the use of a valid test might shed a doubt on a replicability claim inferred from a non-valid test. However, a difficulty in evaluating the results of valid tests is the aforementioned requirement of larger samples. An insignificant finding

[5]Other approaches followed by Camerer et al. (2016) include using the confidence interval of the replication study to see if it contains the effect size of the original study (Cumming, 2008) and using a combining the data sets of both studies to obtain a meta-analytical treatment effect. We comment on the accuracy of those approaches below.

generated by a valid test can be due to insignifficant data, when the study was designed for inference under a non-valid test.

We introduce new terminology to accommodate for the difficulty of making inference when no significant evidence is obtained with the stochastic inequality test. First of all, whether a study replicates or not can only be assessed when the original study exhibits a significant treatment effect using a valid test. We say that the results *strongly replicate* if the replication data also shows a significant treatment effect in the same direction as the original study. On the other hand, we say that the results *weakly replicate* if the replication data does not show a significant opposite treatment effect. When the results weakly replicate then more evidence is needed to better confirm the original finding. Thus, it is only if the study does not weakly replicate then the unique conclusion is that there is doubt as to how representative the original result is. On the other hand, under weak replicability without strong replicability more evidence has to be gathered in order to substantiate the original claim. Only under strong replication the orginal findings are found to robust.

### 2.3.1 Power

Some remarks are in place about the use of power in the traditional replication methodology. It is common practice to cast doubt about the original study if the replicated study shows no significant treatment effect. That is because the sample size of the replication study is chosen so that a rejection is likely if the same effect occurs. There are however several problems with this procedure. First of all, the sample sizes are typically chosen on the basis of non-valid tests. Second, estimates are taken to make inference. The right approach is to use confidence intervals. Third of all, the likelihood of discovering a treatment effect making a wrong conclusion is typically chosen to be 90 % and not 95 %, which makes the false postive equal to 10 % and not 5 %.

The sample size of a replication study should be such that the rejection probability of a valid test is at least 0.95 for all parameters that belong to the 95 % confidence interval of this test. However, a simpler approach is to choose the same sample size of the original study. After all, the original study features a sample size that was designed to uncover a significant treatment effect. So as long as the replication study sample is as large then one can draw some conclusions when the treatment effect is non-significant. The conclusion when observing weak but not strong is that is that the original finding was not as pronounced as it seemed and that more evidence is needed.

### 2.3.2 Magnitude

The magnitude of a treatment effect is an integral part of any study that discovers a treatment effect. Treatment effects of importance should substantiate in similar environments, yet it is not obvious that the magnitudes should be remain identical.

Note that the power analysis used in the traditional replication methodology assumes that the magnitude will be the same in the replication study. This is because the sample sizes are designed based on the effect size in the original study and lack of significance in the replication study is interpreted as lack of replicability.

It is hard to argue that magnitudes have to stay unchanged. However, when magnitudes are allowed to differ between studies then the traditional approach to choosing sample sizes in the replication study can no longer be taken. In fact, it is difficult to specify sample sizes when magnitudes can change. It therefore seems to be a good practice to not change the sample size in the replication study. Afterall, the original study claims that the effect substantiates with the number of observations chosen.

Change in magnitudes should not raise doubt about the validity of the original study. Yet how magnitudes differ helps understanding the treatment effect. Empowered by the stochastic inequality test and its associated confidence intervals for the stochastic tendency we can compare magnitudes. Note that we are not interested in comparing estimates but in comparing the true underlying stochastic tendencies. When the confidence intervals of the stochastic tendencies of the original data and the replicated data do not overlap then we have evidence that the magnitudes are different. The level of significance behind this evidence is easily computed. Following the Šidák (1967) correction, to obtain significant evidence at 5% that the two magnitudes are different one has to consider confidence intervals with confidence levels $1-0.025321 = 0.974679$. We then say that the magnitude (of the original study) replicates if the confidence intervals with this coverage of the two studies overlap. Otherwise we say that the magnitude does not replicate.

When the magnitude replicates the reader gains information about similarity in the findings. When they do not replicate there is evidence of a heterogeneous treatment effect. Evidence of a heterogeneous effect need not shed doubt on the validity and generality of the original study. In particular, when there is evidence that the stochastic tendency in the replication study is larger then there is good news as the effect is even more pronounced. Even when the effect in the replication study is smaller, as long as it is still economically meaningful, no shadow is cast on the original study. Of course, it is a subjective matter whether or not an effect is economically meaningful. In this paper we refrain from such judgements.

### 2.3.3 Meta-Analytic Treatment Effect

One is tempted to combine the two data sets into a single one, to investigate significance in this larger set and then to make claims about replication. This combination would only be valid if one assumes that the data generating processes underlying the two studies are identical. However, this is not a very plausible assumption. To investigate both data sets together requires the use of a valid test that can accommodate for heterogeneity. Such tests exist. However, to incorporate insights based on these tests is outside the scope of this paper.

# 3 Replication Analysis

In this section we apply the methodology presented above to the studies featured in Camerer et al. (2016). We refer to the studies featured therein using the number corresponding to their descending alphabetical order. In Table 1 we provide information about each of the studies such as variable of interest, sample size and range of the variable of interest. Of the 18 studies in Camerer et al. (2016) we have data on 16 of them.[6] The list of the papers that correspond to the study numbers can be found in Table 9 in the Online Appendix.

Our analysis is simplest for studies in which inference is based on pairwise testing, namely studies 3, 6, 7, 8, 9, 11, 13, and 16. These papers use either the t-test or the WMW test. All we have to do is to replace their test with the stochastic inequality test. We follow the convention in the statistical literature and set the level of our tests equal to 5% and the coverage of our confidence intervals analogously to 95%.

The remaining studies use a regression analysis to perform inference. The main disadvantage of using regressions to make inference is similar to that of the t-test explained in the previous section: excessive reliance on distributional assumptions that cannot be readily assumed with a small sample. However, we must emphasize that the pairwise comparison performed by the stochastic inequality test does not map one to one to the regression methodology. That is because a regression is more flexible. It can accommodate correlation between observations by clustering and independent variables can be controlled for (in a linear functional form).

In each of these studies we drop the regression covariates. For three studies (1, 2 and 15) we instead use a pairwise test proposed by the authors in the original paper and that tests the same hypothesis as the regression. For the remaining studies, we average the observations at the relevant level, i.e. the level at which the standard errors are clustered in the regression, and run a pairwise test between treatment and control. Details are given in Table 8 in the Appendix.

[6] We did not have access to the original data of studies 4 and 18. The authors of study 4 did not have the data available and the replication study was designed around the reported statistics in the original paper. The authors of Study 18 did not reply to our e-mails requesting access to their data.

We refer to this way of testing the original hypothesis as the *unconditional investigation*.

To evaluate whether the original finding can be supported in the unconditional investigation, we first run a pairwise t-test. The results are presented in Table 6. Note that the t-test is the analogous test to the linear regression. In five out of the eight studies the t-test reports a significant treatment effect, these are studies 1, 5, 10, 15 and 17). We also find that the WMW test reveals a significant effect (see Table 7 ). So for these five studies the traditional methodology uncovers a treatment effect.

Table 1: Description of Original Studies in Camerer et al. (2016)

| **Study Number** | **Variable of Interest** | **Variable range** | **Method of inference** | **N** |
|---|---|---|---|---|
| 1 | Earnings | [0, 25] | Regression | 120 |
| 2 | Contributions | [0,20] | Regression | 117 |
| 3 | Surplus | [6,23] | WMW test | 216 |
| 5 | Effort | [110,170] | Regression | 72 |
| 6 | Payoff | [0, 1.75] | t-test | 158 |
| 7 | Efficiency Ratio | [0,1] | WMW test | 54 |
| 8 | Efficiency | [0,1] | WMW test | 168 |
| 9 | Gap WTA-WTP | [0,10] | t-test | 112 |
| 10 | Avg. Delegation | [0,1] | Regression | 60 |
| 11 | Median Cooperation rate | [0,1] | WMW test | 78 |
| 12 | Average Cooperation rate | [0,1] | Regression | 124 |
| 13 | Worker Earnings | [0,120] | WMW test | 120 |
| 14 | Discounting | [0,1] | Regression | 69 |
| 15 | Donation average | [0,1] | Regression | 288 |
| 16 | RAD Fundamental Value | [0,1] | WMW test | 120 |
| 17 | Coordination | [1,4] | Regression | 126 |

## 3.1 Treatment Effects in the Original Data Sets

Consider first the eight original studies that are built on pairwise testing. Therein the authors used statistical tests that had more restrictive null hypotheses. The WMW test assumes identical distributions, the two sample t-test assumes normal distribution with equal means. Nevertheless, the stochastic inequality test is able to reject the null hypothesis in six out of the eight studies. In these cases it validates the findings of the original paper. Table 2 presents the results.

Studies 9 and 16 do not exhibit a significant effect. Our interpretation for these findings is as follows. Study 9 employs the t-test which assumes normality. Rejecting the null hypothesis using the t-test means that there is either evidence that the means are different or there is evidence that the data is not normally distributed with equal variances. In this sense, the t-test

does not uniquely identify a treatment effect. Seen a bit differently, the t-test only rejects the null hypothesis if the data is normally distributed. So the original study cannot provide any conclusion unless the data is normally distributed. Attempting to uncover an effect without assuming normal distribution by applying the stochastic inequality test fails. The valid nature of the stochastic inequality test allows us to make statistical inference free of these assumptions. Study 16 uses the WMW test and finds evidence to reject the null hypothesis of no difference in distributions between treatment and control. The fact that the stochastic inequality does not reject the null hypothesis, indicates that the original study is uncovering changes in higher moments of the distributions of treatment and control and not changes in location.

Table 2: Stochastic Inequality Applied to Original Studies in Camerer et al. (2016)

| **Study Number** | **Stochastic Inequality C.I. (95%)** | **Stochastic Tendency** $d$ | **Replicates in Camerer et al (2016)?** |
|---|---|---|---|
| **Pairwise test studies** | | | |
| 3 | [2.1,8.4] | $5.9^{**}$ | Yes |
| 6 | [0.908,0.91] | $0.909^{***}$ | Yes |
| 7 | [-0.26, -0.08] | $-0.16^{***}$ | No |
| 8 | [0.41,0.86] | $0.66^{***}$ | Yes |
| 9 | [-0.4,2] | 0.89 | No |
| 11 | [0.33,0.52] | $0.42^{***}$ | Yes |
| 13 | [-83, -27] | $-52^{**}$ | No |
| 16 | [-0.72,0.01] | $-0.33^{*}$ | Yes |
| **Regression studies** | | | |
| 1 | [ -0.1, 4] | $1.85^{*}$ | No |
| 2 | [-11.2,2.7] | -2.8 | Yes |
| 5 | [13, 38] | $23^{***}$ | No |
| 10 | [0.04, 0.40] | $0.20^{**}$ | Yes |
| 12 | [-0.25, 0.13] | -0.09 | Yes |
| 14 | [-0.06, 0.20] | 0.08 | No |
| 15 | [-0.04, 0.24] | 0.09 | Yes |
| 17 | [-1.4, 0.01] | -0.57 | Yes |

Notes: This table presents the stochastic tendency $d$ of the original studies (col. 3) and 95% confidence interval of the stochastic inequality test (col. 2). *** denotes significance at the 1% level, ** denotes significance at the 5% level, and * denotes significance at the 10% level.

Consider now the eight remaining studies that employ a regression analysis in the original study. The stochastic inequality test is only able to uncover a significant treatment effect in two of the unconditional versions of the original studies, namely 5 and 10. In the remaining studies there is no significant evidence of a treatment effect. Thus, in three cases, studies 1, 15, and 17,

the stochastic inequality test does not confirm the findings of the t-test.

## 3.2 Treatment Effects in the Replication Studies

The sample size for each replication study was collected so that, given the test performed in the original study and the effect size found there, the probability that the test correctly rejects the null hypothesis is, at least, 0.9. In six out the eight replication pairwise studies the sample sizes of the replication data were considerably larger than that of the original study. In two out of the eight studies based on regressions sample sizes in the replication data were considerably smaller (these were studies 15 and 17) and in other two studies the samples across replication and original data sets were similar (studies 4 and 12).

We evaluate the treatment effects in the replication data sets. At this point we treat each data set as a self contained study, and use the stochastic inequality test to evaluate treatment effects. As in the previous section, we first consider the studies that used pairwise testing. Table 3 presents the effect sizes and confidence intervals for each of those studies. In five out of the eight studies the replication study reveals a significant treatment effect. Only studies 7, 9 and 13 do not reveal a significant treatment effect in the replication data.

We also apply the stochastic inequality test to each of the studies that used regressions to make inference. We find that studies 2, 12 and 15 exhibit a significant treatment effect. The remainder of studies do not exhibit a treatment effect.

## 3.3 Treatment Effects Overall

We find that the stochastic inequality test uncovers treatment effects in 15 out of the 21 studies (71 %) for which the conventional treatment effect uncovers one. For the original data sets it uncovers a significant treatment effect in eight out of 13 studies (61,1 %) and for the replication data sets it uncovers a significant treatment effect in seven out eight studies (87%). This shows that while the test is more stringent, as it does not rely on distributional assumptions and tests a more general hypothesis, it is powerful. It is able to reject the null in most of the instances in which the conventional test does so.

## 3.4 Replication Analysis

### 3.4.1 Standard Approach

The most widespread way to establish the replicability of findings is to evaluate whether the replication data set has a significant effect in the same direction as the original data set (Nosek and Lakens, 2014; OSC, 2015; Camerer et al., 2016). As discussed in section 3.2, we refer to this

Table 3: Stochastic Inequality Applied to Replication Studies in Camerer et al. (2016)

| **Study Number** | **N** | **Stochastic Inequality C.I. (95%)** | **Stochastic Tendency** $d$ | **Treatment Replicates?** | **Replicates in Camerer et al. (2016)?** |
|---|---|---|---|---|---|
| **Pairwise test studies** | | | | | |
| 3 | 360 | [1.80,12.2] | $7.1^{***}$ | Strong | Yes |
| 6 | 156 | [1.66,1.68] | $1.67^{***}$ | Strong | Yes |
| 7 | 96 | [-0.12, 0.20] | 0.026 | Weak | No |
| 8 | 128 | [0.33, 0.86] | $0.61^{***}$ | Strong | Yes |
| 9 | 250 | [-0.20,1.40] | 0.64 | NA | No |
| 11 | 40 | [0.13, 0.15] | $0.147^{***}$ | Strong | Yes |
| 13 | 160 | [-10, 40] | 13.7 | Weak | No |
| 16 | 220 | [-0.23,-0.01] | $-0.11^{**}$ | Weak$^{r}$ | Yes |
| **Regression studies** | | | | | |
| 1 | 318 | [ -0.2, 2.1] | 0.66 | NA | No |
| 2 | 340 | [-8.5,-1.4] | $-4.8^{***}$ | Weak$^{r}$ | Yes |
| 5 | 168 | [-1.4,12] | 5.2 | Weak | No |
| 10 | 102 | [-0.09, 0.57] | 0.21 | Weak | Yes |
| 12 | 120 | [-0.26, -0.01] | $-0.165^{***}$ | Weak$^{r}$ | Yes |
| 14 | 131 | [-0.06, 0.06] | 0.005 | NA | No |
| 15 | 48 | [-0.01, 0.53] | $0.24^{*}$ | NA | Yes |
| 17 | 90 | [-1.76, 0.25 ] | -0.71 | NA | Yes |

Notes: This table presents the stochastic tendency $d$ of the replication studies (col. 3) and 95% confidence interval of the stochastic inequality test (col. 2). In column 5 "Strong" denotes strong replicability (as defined in Sec. 2.3), " Weak" denotes weak replicability (as defined in Sec. 2.3), and "NA" denotes not applies since there was not a treatment effect in the original study to begin with. *** denotes significance at the 1% level, ** denotes significance at the 5% level, and * denotes significance at the 10% level. $r$ denotes that the conclusion of weak replicability is reached when the role of the original and replication study are reversed.

as replication in the strong sense. Accordingly, we find that four out of the eight studies strongly replicate when a valid test is used (Tables 2 and 3). These are studies 3, 6, 8, and 11.

Furthermore, we find that studies 7 and 13 replicate in the weak sense. We differ from Camerer et al. (2016) in how we interpret the finding that the replication data does not exhibit a treatment effect while the corresponding original data set does. We categorize situations in which the replicated data *does not contradict the original findings* as replication in the weak sense. This applies to these two studies. Unlike Camerer et al. (2016) who casts doubt about the validity of the original findings using a non-valid test, our interpretation is that not sufficient data has been gathered. Consider now the two pairwise-test studies for which we cannot identify a significant effect in the original data set with a valid test (9 and 16). In study 16 the treatment effect uncovered by the stochastic inequality test in the original data is marginally

significant, while it is significant in the replication data. It seems like too little data was originally gathered (120 in the original data versus 220 in the replicated data). On the other hand, study 9 has an insignificant treatment effect in both the original and replicated data sets. Here the treatment effect, if any, seems less pronounced than originally noted.

Next, consider the unconditional investigations corresponding to the studies that used a regression analysis. None of these studies strongly replicate as we find no significant treatment effects in both original and replication data. Morerover, the two studies that exhibit a treatment effect in the original data (5 and 10) fail to show any significant result in their corresponding replication data. This result emerges despite the fact that these data sets have a larger sample size. According to our terminology, these studies replicate in the weak sense. On the one hand, more data is needed to establish whether the original claims can be established with our unconditional investigation. On the other hand, there is some doubt that the claims from the original study are so general.[7]

There are two studies that exhibit a significant treatment effect in the replication data (studies 2 and 12) but insigificance in the original data set. We obtain this finding not only with the stochastic inequality test but also with the t-test (Table 6). In study 2 the replication data set was much larger, hence pointing to lack of data in the original data set. The sample sizes in study 12 were very similar, so a potential explanation is that the treatment effect is much different between data sets. If we were to swap the role of original and replicated study we would be noting that these two studies replicate in the weak sense. This leads us to highlight them in Table 3 with an "$r$".

The remaining four studies (1, 14, 15 and 17) show insignificance in the original and replication study. In particular, they do not qualify for a replication analysis. Yet three of these studies showed significance in the original data set when a conventional test was used (1, 15, 17). Also, study 14 is borderline significant under the conventional test. Hence, it is not that our unconditional analysis impedes observing treatment effects. Instead, the treatment effects uncovered by the t-test are too dependent on the distributional assumptions of that test, and for that reason they are insignificant under the stochastic inequality test. This strong dependence on distributional assumptions lead the replication team to gather less data for the replication data of studies 15 and 17, making it more difficult to observe any treatment effect when a valid test is used. It is interesting to note that study 15 and 17 replicate according to the standards of Camerer et al. (2016).

[7]The interpretation is similar for study 1 where the stochastic inequality applied to the original data gives a marginally significant effect but the replicated data, which is considerably larger, gives insignificance.

### 3.4.2 Treatment and Magnitude Replication

We now combine the concept of "treatment replicates" with the concept of " magnitudes replicates" to obtain a richer standard for replicability. In table 4 we present the relevant confidence intervals for each data set. Column 4 in that table indicates whether, for a study with a significant treatment effect, its magnitude replicates. As defined in Section 2.3., this means that the confidence intervals of original and replication data sets, presented in columns 2 and 3, respectively, overlap.

We observe the strongest case for replication in studies 3 and 8. Both studies show significant treatment effects in the original data set and the confidence intervals of original and replication data sets overlap. We speak of replication of the treatment effect in the strong sense together with replication of the magnitude. For study 6, there is replication of the treatment in the strong sense but the magnitude of the treatment effect in the replication data is larger. While in this case the magnitude of the treatment effect does not replicate, this stronger treatment effect is desirable. Study 11 exhibits the opposite case. There is replication of treatment in the strong sense but no replication of the magnitude, the replication is showing a weaker treatment effect.

For study 7 the replication data set does not add substantial information. There is no treatment effect in the replication data and the confidence intervals of both data sets overlap. Consequently, we have replication of the treatment effect in the weak sense and replication of the magnitude. The same conclusion holds for study 16 if the roles of original and replication data sets are reversed. Finally, in study 13 the replication data shows lower magnitude without a significant treatment effect. Hence, there is replication of the treatment effect in the weak but not in the strong sense and no replication of the magnitude. It seems as if the treatment effects of original and replication data emerge from different experimental designs. Note that study 9 has no significant treatment effect in either study and hence cannot be investigated. So we conclude that for the magnitude replicates in four out of the seven pairwise-test studies with a significant effect in either data set.

We can make inference for four out of the eight unconditional investigations based on the studies using regressions. That is because, as we mentioned in the previous section, there are four studies in which neither original nor replication data exhibit a significant treatment effect. Thus, our focus is on studies 2, 5, 10, and 12. We find that for all four cases the magnitude replicates. Studies 5 and 10 exhibit a treatment effect in the original dataset. For both studies, the corresponding replication study adds no information as the treatment effect does not exhibit significance. Moreover, we highlight that for study 5 the overlap of confidence intervals is rather thin and would not have overlapped if we had considered evidence at the 5% level. For these studies there is evidence of replication of the treatment effect in the weak sense together with

Table 4: Confidence Intervals for Comparing Treatment Effects of the Studies in Camerer et al. (2016)

| **Study** | **Stochastic Inequality 97.47% C.I. Original Study** | **Stochastic Inequality 97.47% C.I. Replication Study** | **Magnitude Replicates?** | **Replicates in Camerer et al. (2016)?** |
|---|---|---|---|---|
| **Pairwise test studies** | | | | |
| 3 | [1, 9.4] | [1.6, 12.8] | Yes | Yes |
| 6 | [0.0908, 0.91] | [0.166, 0.168] | No | Yes |
| 7 | [0.06, 0.29] | [-0.12, 0.21] | Yes | No |
| 8 | [0.38, 0.88] | [0.31, 0.92] | Yes | Yes |
| 9 | [-0.3, 2] | [-0.4, 1.5] | NA | No |
| 11 | [0.32, 0.56] | [0.13, 0.15] | No | Yes |
| 13 | [-86, -26] | [-16, 44] | No | No |
| 16 | [-0.74, 0.11] | [-0.28, -0.01] | Yes | Yes |
| **Regression studies** | | | | |
| 1 | [-0.3, 4.1] | [-0.3, 2.3] | NA | No |
| 2 | [-11.5, 3.1] | [-9.1,-1] | Yes | Yes |
| 5 | [11, 37] | [-2.2, 12] | Yes | No |
| 10 | [0.01, 0.43] | [-0.13, 0.56] | Yes | Yes |
| 12 | [-0.25, 0.13] | [-0.37, -0.01] | Yes | Yes |
| 14 | [-0.07, 0.21] | [-0.08, 0.07] | NA | No |
| 15 | [-0.05, 0.25] | [-0.07,0.54] | NA | Yes |
| 17 | [-1.5, 0.12] | [-1.87, 0.25] | NA | Yes |

Notes: This table presents the confidence interval of the stochastic inequality test with Sidak correction applied to original studies (col. 2) and replication studies (col. 3). In column 5 "Yes" indicates that the confidence intervals of the original and replication studies overlap and the stochastic inequality reveals a significant treatment effect in the original study, "No" indicates that the confidence intervals of original and replication studies do not overlap and the stochastic inequality reveals a significant treatment effect in the original study, "NA" indicates that the stochastic inequality did not reveal a treatment effect in the original study.

replication of the magnitude. The same conclusion holds for studies 2 and 12 when the roles of replication and original data are reversed.

# 4 Comparing replication studies

We further discuss our results and compare them to those of Camerer et al. (2016). To facilitate reading, we first summarize our findings. Table 5 compiles the assessment of significance for each study and each data set according to the stochastic inequality test and the pairwise test used in the original study. It also includes the replication assessments of Camerer et al. (2016).

Table 5: Summary Table

| | **Original Study** | | **Replication Study** | | **Conclusions** | | |
|---|---|---|---|---|---|---|---|
| **Study** | **Significance t-test or MWM?** | **Significance Stochastic Inequality?** | **Significance t-test or MWM?** | **Significance Stochastic Inequality?** | **Treatment Replicates?** | **Magnitude Replicates?** | **Replicates according to Camerer et al. (2016)?** |
| **Pairwise Test Study** | | | | | | | |
| **3** | Yes | Yes | Yes | Yes | Strong | Yes | Yes |
| **6** | Yes | Yes | Yes | Yes | Strong | No | Yes |
| **7** | Yes | Yes | No | No | Weak | Yes | No |
| **8** | Yes | Yes | Yes | Yes | Strong | Yes | Yes |
| **9** | Yes | No | Yes | No | NA | NA | No |
| **11** | Yes | Yes | Yes | Yes | Strong | No | Yes |
| **13** | Yes | Yes | No | No | Weak | No | No |
| **16** | Yes | No | Yes | Yes | Weak$^{r}$ | Yes | Yes |
| **Regression Study** | | | | | | | |
| **1** | Yes | No$^{m}$ | No$^{m}$ | No | NA | NA | No |
| **2** | No | No | Yes | Yes | Weak$^{r}$ | Yes | Yes |
| **5** | Yes | Yes | No$^{m}$ | No | Weak | Yes | No |
| **10** | Yes | Yes | Yes | No | Weak | Yes | Yes |
| **12** | No | No | Yes | Yes | Weak$^{r}$ | Yes | Yes |
| **14** | No$^{m}$ | No | No | No | NA | NA | No |
| **15** | Yes | No | Yes | No | NA | NA | Yes |
| **17** | Yes | No$^{m}$ | No$^{m}$ | No | NA | NA | Yes |

Notes: This table summarizes the findings of the paper. Significance of a treatment effect a the 5% level as assessed by standard pairwise testing for original studies is given in Col. 2 and for replication studies in Col. 4. Significance of a treatment effect a the 5% level as assessed by the stochastic inequality test for original studies is given in Col. 3 and for replication studies in Col. 4. Column 6 presents information about treatment replicability, as defined in Sec. 2.3. Column 7 presents information about magnitude replicability as defined in Sec. 2.3. $r$ denotes that the conclusion of weak replicability is reached when the role of the original and replication study are reversed. $m$ denotes that the conclusion is reversed when the significance level is relaxed and set at 10.%

There are four studies that replicate in the strong sense according to the stochastic inequality test, namely studies 3, 6, 8, and 11. Note that these studies also replicate in Camerer et al.

(2016). Hence, our strongest assessment for treatment effects, "strong replicability", coincides with the assessment of Camerer et al. (2016) for four out of the 10 studies that replicate according to their standards. Importantly, our methodology enables us to talk about magnitudes over and beyond treatment replicability. The magnitude of these four studies is replicated in two studies (3 and 8). For study 6, the treatment effect in the replication study is larger than in the original study while for study 11 the opposite is true.

There are three studies, 1, 9, and 14, for which the stochastic inequality could not detect a treatment effect in the original data set. These studies did not replicate according to Camerer et al. (2016). This constitutes 50% of the studies that did not replicate according to their standards. Our conclusion is that the test performed to make inference in the original study, and also used to design the replication study, revealed a treatment effect that was not there in this extent to begin with.

We question the positive assesment of Camerer et al. (2016) for studies 15 and 17. Under the stochastic inequality test they do not exhibit a significant treatment effect in the original study. This is not an artifact of our data manipulation to fit pairwise testing, as shown by the significant treatment effects found for these data sets in Tables 6 and 7. Instead, this null result stems from the assumption-free nature of the stochastic inequality test vis-a-vis the t-test. In our view their conclusion is guided by invalid testing.

There are seven studies for which we find weak replication. These are studies 2, 5, 7, 10, 12, 13, and 16. Rather than claiming that their claims are invalidated, we first look at their replicability of magnitude to obtain more information. We find that the magnitude of study 13 cannot be replicated, pointing at a considerable heterogeneity in treatment effects between replication and original data. For study 5 we note that the confidence intervals of original and replication data sets intersect on a thin set. On an intuitive level there seems to be evidence of heterogenity in the treatment effect for this study. However, formally it cannot be ruled out that the treatment effect lies in this thin intersection of the intervals. These two studies were found not to replicate by Camerer et al. (2016). A stark difference in magnitude of the treatment across original and replication experiments might be the reason for that conclusion. We note that such difference in treatment effects is as if original and replication data sets are generated from different experimental designs. This explanation that coincides with the assessment of Chen et al. (2021).

For the remainder of studies with weak replication, we do find compelling evidence for replication of magnitudes. Recall that the stochastic inequality test requires larger sample sizes, as it is free of distributional assumptions and evaluates tendencies. Hence, we conclude that more data is needed to better understand the treatment effects in these studies. Therefore, the assessment of Camerer et al. (2016) that study 7 does not replicate is assessed to be premature

and relies on invalid testing.

Finally, we note that there are three studies for which the stochastic inequality test uncovered a significant treatment effect in the replication study but not in the original study. These are studies 2, 12, and 16, all of which replicate according to Camerer et al. (2016). The finding that the replication data yields a significant treatment effect whilethe original does not also holds for studies 2 and 12 when the conventional t-test and WMW test are used (see Tables 6 and 7). Instead for study 16, the MWM test gives a significant treatment effect in both data sets.

Our conclusion differs per study. In study 2, the larger data size in the replication data facilitated observing a significant treatment effect for both the stochastic inequality and the t-test. This is an instance in which the findings of the regression analysis happen to extend to pairwise testing with a valid test given the larger data set. A similar conclusion is found for study 16. The larger sample size in the replication data allowed the stochastic inequality to detect the significant treatment effect that was uncovered by the MWM test. In this case the differences in distribution captured by the latter test translated into differences in tendencies as captured by the former. In study 12 original and replication data sets have a similar size. Hence, it is unknown which finding is more informative. Our conclusion is that the robustness of these findings in light of our test can only be achieved with a low number of observations.

# 5 Conclusion

Replicability of laboratory experiments is of central importance. We introduce a novel methodology for drawing conclusions from replication studies. One innovation is that we use the stochastic inequality test. This is a non-parametric and valid test for uncovering treatment effects. An additional innovation is that we propose how to evaluate evidence in replication studies. In particular, we argue why inference from insignificant findings cannot be made when power calculations are based on estimates.

We apply our methodology to the studies investigated by Camerer et al. (2016) We confirm the replicability in four of the studies. In seven studies this new test can only partially uncover treatment effects, possibly indicating that samples sizes are not large enough. In particular we differ here from the convention to claim lack of replicability if there is no significant effect in the replication study. On the side our findings are reassuring about the power of the stochastic inequality test. It roughly uncovers a treatment in roughly 71% of the cases where the conventional t-test and WMW test detect a treatment effect.

# References


**Agresti, Alan**, *Analysis of ordinal categorical data*, New York: Wiley, 1984.

**Bahadur, Raghu R and Leonard J. Savage**, "The nonexistence of certain statistical procedures in nonparametric problems," *The Annals of Mathematical Statistics*, 1956, *27* (4), 1115–1122.

**Brunner, Edgar and Ullrich Munzel**, "The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 2000, *42* (1), 17–25.

**Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan et al.**, "Evaluating replicability of laboratory experiments in economics," *Science*, 2016, *351* (6280), 1433–1436.

**Charness, Gary and Peter Kuhn**, "Lab labor: What can labor economists learn from the lab?," *Handbook of labor economics*, 2011, *4*, 229–330.

**Chen, Yan, Peter Cramton, John A. List, and Axel Ockenfels**, "Market design, human behavior, and management," *Management Science*, 2021, *67* (9), 5317–5348.

**Cliff, Norman**, "Dominance statistics: Ordinal analyses to answer ordinal questions," *Psychological bulletin*, 1993, *114* (3), 494.

**Cumming, Geoff**, "Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better," *Perspectives on psychological science*, 2008, *3* (4), 286–300.

**Dechenaux, Emmanuel, Dan Kovenock, and Roman M. Sheremeta**, "A survey of experimental research on contests, all-pay auctions and tournaments," *Experimental Economics*, 2015, *18* (4), 609–669.

**Duffy, John**, *Experiments in Macroeconomics*, Emerald Group Publishing, 2014.

**Hamermesh, Daniel S.**, "Replication in economics," *Canadian Journal of Economics/Revue canadienne d'économique*, 2007, *40* (3), 715–733.

**Kagel, John H. and Dan Levin**, "Auctions: A survey of experimental research, 1995-2010," *Handbook of experimental economics*, 2011, *2*, 563–637.

**Lehmann, Erich L. and Wei-Yin Loh**, "Pointwise versus uniform robustness of some large sample tests and confidence intervals," *Scandinavian Journal Statististics*, 2000, *17*, 177–187.

**Mann, Henry B and Donald R. Whitney**, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, 1947, pp. 50–60.

**Nosek, Brian A. and Daniël Lakens**, "Registered reports," *Social Psychology*, 2014.

**OSC, (Open Science Collaboration)**, "Estimating the reproducibility of psychological science," *Science*, 2015, *349* (6251).

**Schlag, Karl H.**, "A New Method for Constructing Exact Tests without Making any Assumptions," *Department of Economics and Business Working Paper 1109, Universitat Pompeu Fabra*, 2008.

**Šidák, Zbyněk**, "Rectangular confidence regions for the means of multivariate normal distributions," *Journal of the American Statistical Association*, 1967, *62* (318), 626–633.

**Student**, "The probable error of a mean," *Biometrika*, 1908, pp. 1–25.

**Wilcoxon, Frank**, "Individual comparisons by ranking methods," 1945, pp. 80–83.

# Appendix

## Replication Analysis of Camerer et al. (2016) with other Pairwise Tests

Table 6: Replication Analysis of Camerer et al. (2016) using t-tests

| **Study** | **t-test original** | **95% C.I. original** | **t-test replication** | **95% C.I. replication** | **Replicates in Camerer et al. 2016)?** |
|---|---|---|---|---|---|
| **Pairwise test studies** | | | | | |
| 3 | 5.826*** | [3.658, 8.188] | 4.585*** | [3.877, 10.433] | Yes |
| 6 | 4.776*** | [0.045, 0.109] | 7.713*** | [0.124, 0.209] | Yes |
| 7 | -2.855** | [-0.286,-0.026] | 0.529 | [-0.079, 0.131] | No |
| 8 | 8.967*** | [0.504, 0.811] | 6.737*** | [0.418, 0.809] | Yes |
| 9 | 2.260** | [0.038,0.595] | 1.931* | [-0.004, 0.443] | No |
| 11 | 24.043*** | [0.385, 0.453] | 5.582*** | [0.095, 0.199] | Yes |
| 13 | -6.435*** | [-69.923, -33.957] | 1.780* | [-2.806, 30.214] | No |
| 16 | -3.492*** | [ -0.548, -0.121] | -1.723* | [-0.239, 0.023] | Yes |
| **Regression studies** | | | | | |
| 1 | 2.017** | [0.0342, 3.665] | 1.406 | [0.265,1.598] | No |
| 2 | -1.279 | [-7.233, 1.633] | -4.632*** | [-6.849, -2.747] | Yes |
| 5 | 7.183*** | [17.229, 30.474] | 1.737* | [-0.711,11.120] | No |
| 10 | 3.221*** | [0.076, 0.331] | 2.149** | [0.014, 0.417] | Yes |
| 12 | -1.490 | [-0.211,0.030] | -3.413*** | [ -0.254,-0.067] | Yes |
| 14 | 1.734* | [-0.012, 0.166] | 0.092 | [-0.064, 0.071] | No |
| 15 | 1.956** | [0.001, 0.177] | 2.462** | [0.043, 0.434] | Yes |
| 17 | -2.201** | [-1.112, -0.038] | -2.038* | [-1.438, 0.022] | Yes |

Notes : This table presents t-tests applied to original studies (col. 2) and replication studies (col. 4). It also presents the 95% confidence interval of original studies (col. 3) and replication studies (col. 5). *** denotes significance at the 1% level, ** denotes significance at the 5% level, and * denotes significance at the 10% level.

Table 7: Replication Analysis of Camerer et al. (2016) using the MWM test

| **Study** | **WMW test original** | **WMW Replication** | **Replicates in Camerer et al. 2016)?** |
|---|---|---|---|
| **Pairwise test studies** | | | |
| 3 | $2.722^{***}$ | $3.250^{***}$ | Yes |
| 6 | $4.748^{***\rho}$ | $8.424^{***\rho}$ | Yes |
| 7 | $-2.449^{**}$ | 0.420 | No |
| 8 | $3.873^{***}$ | $3.361^{***}$ | Yes |
| 9 | $1.675^{*}$ | $1.647^{*}$ | No |
| 11 | $21.885^{***}$ | $7.730^{***}$ | Yes |
| 13 | $-2.882^{***}$ | 1.470 | No |
| 16 | $-2.402^{**}$ | $-2.594^{***}$ | Yes |
| **Regression studies** | | | |
| 1 | $2.426^{**}$ | $1.881^{*}$ | No |
| 2 | -1.476 | $-3.735^{***}$ | Yes |
| 5 | $5.021^{***}$ | $1.950^{*}$ | No |
| 10 | $3.029^{***}$ | $1.831^{*}$ | Yes |
| 12 | 1.405 | $-3.128^{***}$ | Yes |
| 14 | 1.219 | 0.318 | No |
| 15 | $2.378^{**}$ | $2.359^{**}$ | Yes |
| 17 | $-2.160^{**}$ | -1.633 | Yes |

Notes : This table presents MWM applied to original studies (col. 2) and replication studies (col. 3). *** denotes significance at the 1% level, ** denotes significance at the 5% level, and * denotes significance at the 10% level. $\rho$ denotes that a Wilcoxon signed-rank test was performed instead of a WMW due to the matched nature of the data.

## A2. Description of analyses and data transformation

Table 8: Description of analyses and data transformations

| **Study Number** | **Original analysis to evaluate treatment effects** | **Alternative pairwise analysis in paper** | **Proposed data transformation to fit pairwise testing** |
|---|---|---|---|
| 1 | OLS regression using effort as dependent variable. | Footnote 11. WMW test | Footnote 11. WMW test |
| 2 | OLS regression clustered at the group level using earnings as dependent variable. | pg. 3322 paragraph 4. WMW test | pg. 3322 paragraph 4. |
| 3 | WMW test using matching group averages as the unit of observation. | | None |
| 5 | Random effects regression clustering both at the session and individual level and with effort as dependent variable. | None | Effort averaged at the individual level |
| 6 | unpaired t-test comparing average aggregate payoffs. | | None |
| 7 | WMW test of efficiency ratios averaged at the session level. | | None |
| 8 | WMW test with matching groups of 8 subjects as one independent observation. | | None |
| 9 | t-test across treatments. | | None |
| 10 | Probit regression controlling for period fixed effects. Standard errors are clustered at the individual level. | None | Contribution averaged at the individual level |
| 11 | WMW test on subject median cooperation rate across all periods. | | None |
| 12 | Logistic regression over all individual decisions clustered on both subject and interaction pair. | | Mean decision by group and subject |
| 13 | Mann-Whitney test comparing average sums of worker earnings across treatments. | | None |
| 14 | OLS regression clustered at the individual level with present value as dependent variable. Subjects who did not discount at all were excluded. | Table 2D last column and last row. t-test. | Table 2D last column and last row. t-test. |
| 15 | OLS regression with clustering at the subject level and robust standard errors and decision of choosing to be a donor as dependent variable. | None | Mean donation at the subject level |
| 16 | WMW test of pairwise comparisons between treatments. | | None |
| 17 | Ordered probit regression with group random effects and Huber-White standard errors. | None | Mean output by session and group |

# Online Appendix

Table 9: Studies included in Camerer et al. (2016) with authors' identity

| **Number** | **Study** | **Journal** |
|---|---|---|
| 1 | Abeler et al. (2011) | AER |
| 2 | Ambrus and Greiner (2012) | AER |
| 3 | Bartling et al. (2012) | AER |
| 4 | Charness and Dufwenberg (2011) | AER |
| 5 | Chen and Chen (2011) | AER |
| 6 | de Clippel et al. (2014) | AER |
| 7 | Duffy and Puzzelo (2014) | AER |
| 8 | Dulleck et al. (2011) | AER |
| 9 | Ericson and Fuster (2011) | QJE |
| 10 | Fehr et al. (2013) | AER |
| 11 | Friedman and Oprea | AER |
| 12 | Fudenberg et al. (2012) | AER |
| 13 | Huck et al. (2011) | AER |
| 14 | Ifcher ad Zarghmee (2011) | AER |
| 15 | Kessler and Roth (2012) | AER |
| 16 | Kirchler et al. (2012) | AER |
| 17 | Kogan et al. (2011) | AER |
| 18 | Kuziemko (2014) | QJE |