universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Statistical reporting and its credibility in student theses of economic psychology: An analysis of 254 theses from 2000-2016"

verfasst von / submitted by

Johanna Kathrin Mosen, B.Sc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (M.Sc.)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Psychologie
Betreut von / Supervisor:	Univ.-Prof. Dr. Erich Kirchler

Eidesstattliche Erklärung

Ich versichere, dass ich die Masterarbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe, und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat. Alle Ausführungen der Arbeit, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Wien, Februar 2017

Unterschrift

Danksagung

An dieser Stelle möchte ich mich bei allen Menschen bedanken,
die mich bei meiner Masterarbeit unterstützt haben.

Ein besonderer Dank gilt Mag. Jerome Olsen,
der nicht nur das Thema dieser Masterarbeit zur Wahl aufgestellt hat,
sondern mich jederzeit bei Fragen unterstützt und viele hilfreiche Inputs geliefert hat.

Außerdem danke ich Mag. Martina Brandtner,
die mir bei der Durchsicht und der Kodierung der Abschlussarbeiten geholfen hat.

Und ich danke Univ.-Prof. Dr. Erich Kirchler, dass er meine Masterarbeit betreut,
unterstützt, ermöglicht und Anreize geliefert hat.

Und schließlich gilt auch meiner Familie, meinem Freund, Freunden und Kommilitonen
besonderer Dank für ihre Geduld und ihre Hilfe während meiner Studienzeit.

Doch der größte Dank gebührt meinen Eltern. Dankeschön, dass ihr mir die
Absolvierung dieses Studium ermöglicht habt und immer hinter mir steht.

Table of content

1. Introduction

In the summer of 2015, psychological science experienced great attention as an international team of 270 researchers reported results of a large-scale reproducibility project. In this project, 100 experiments published in major psychological journals were replicated. It was found that in more than half of the studies the results could not be replicated (Open Science Collaboration, 2015). What are the consequences? In any case, it causes doubts about the scientific enterprise.

Already in 1959, Sterling found evidence that researchers were hoping for significant results in statistical tests, or rather, that they did not publish non-significant results and that significant results are seldom verified by independent replication. Thus, the fact that published results of scientific studies are not a representative sample of all scientific investigations has long been known. No change in this practise has been recorded and hence still exists (Fanelli, 2010; Sterling, Rosenbaum, & Weinkam, 1995). The phenomenon that significant or positive results are reported more frequently and that data in scientific journals is statistically distorted, is called *publication bias* (Sterling, 1959). Furthermore, the term *file drawer problem* is used, which refers to non-significant or negative results remaining in the drawer and not getting published (Rosenthal, 1979).

However, the distorted presentation of results does not provide any explanation for the fact that researchers failed in so many replications - *p-hacking* does. It describes the manipulation of research results, in order to achieve statistical significance (Simonsohn, Nelson, & Simmons, 2014). It is referred to, inter alia, as *p*-hacking, because the statistical significance of a finding is assessed based on an empirical *p*-value, mostly with an alpha level of .05. If the empirical *p*-value is smaller than the alpha level, a finding is regarded as significant. In a research culture where mostly significant results can be published, researchers want to reduce the *p*-value in case of non-significance (Head, Holman, Lanfear, Kahn, & Jennions, 2015). Of course, many errors on the part of the researchers and on the part of the replicators can justify failing in replications, but the results do not show just a small error rate. In fact, the literature shows that scientists distort their statistics in various ways and that they even admit it (John, Loewenstein, & Prelec, 2012).

One reason why researchers manipulate can be found in the established scientific culture, which is called ‘getting it published’ instead of ‘getting it right’

(Asendorpf et al., 2013, p. 114). Scientists know that the probability of getting papers published is greater when results are significant, unusual and interesting, and they know publications can advance a career.

As a result, in the literature the term *questionable research practice* (QRP) is becoming more and more established. QRP can be identified by investigating the statistical reporting, retrospectively the statistical characteristics and values. Hence, the aim of this master thesis is to investigate statistical reporting and its credibility.

Statistics are often the essential supporting structure of a study and their coherent presentation are important for the interpretation of data. Thus, its truthful and consistent presentation plays a vital role for the entire discipline of psychology. Without reliable statistic, no one can do research.

The difference to previous studies is that this thesis examines the credibility of reported statistics among student theses from economic psychology at the University of Vienna. This results in the research question, whether QRPs can also be found among students? If data distortion, like the publication bias or *p*-hacking, is not found in students results, it could mean that students are not influenced by the publication pressure like scientists are. However, if data distortion is found, it would mean that causes for QRPs already lie in the teaching of statistics at the university or in the understanding of the students. The first is expected to be more likely; no publication pressure, no QRPs and therefore no data distortion.

1.1. Disposition

In the following, the theoretical background will be examined first, respectively, some of the results of statistical reporting, that have been found so far. Subsequently, the concrete research questions will be derived from the presented methods and results of the literature. Then the methods and the sample will be explained and the results will be reported and compared with individual findings from the theoretical background part. Within the result part, exploratory results and further tests will also be reported. In addition, based on reporting analyses, a specially developed statistical guide for students will be presented. Finally, the limitations and the discussion will be debated.

2. Theory

2.1. The Hierarchy of Science

The *Hierarchy of Science* (HoS) states that physical sciences are settled at the bottom, biological sciences in the middle, and so-called 'softer sciences', like social sciences, in which the discipline of psychology is found, are located at the top (Comte, 1830). The meaning of this order is that the higher the discipline, the more complex it becomes and the generalizability within the field becomes more difficult. According to this, 'softer sciences' at the top have fewer limitations with regard to errors, which in turn can lead to more positive results compared to disciplines on lower hierarchical levels (Comte, 1830).

Fanelli (2010) verified the Hierarchy of Science using 2434 published papers that engage in hypothesis testing. He showed that the number of 'positive' results increases up the hierarchy (Fanelli, 2010). Figure 1 shows the comparison of the disciplines and that the results confirm the HoS; psychology and psychiatry lead the list with 92% significant or positive results and space science has the fewest significant results, with 70%.

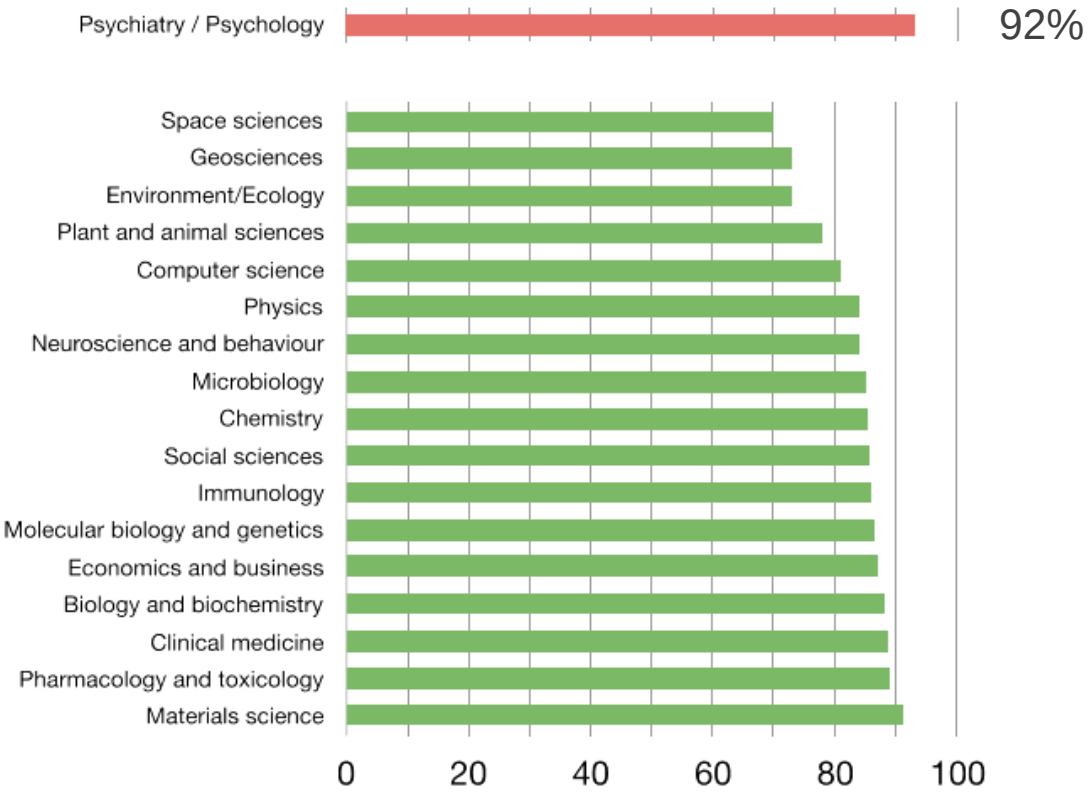

Figure 1. Percentage of papers reporting a support for the tested hypothesis (Fanelli, 2010).

Furthermore, Fanelli found that the frequency of negative results was significantly higher when several hypotheses were tested, in comparison to when only one hypothesis was tested, independently of the disciplines. In addition, the statistical power was higher in disciplines at the bottom such as physics, chemistry or geology (Fanelli, 2010).

The findings show that psychology and psychiatry have the highest number of significant results, yet not the highest statistical power. This missing link makes one suspect that there is possibly an existing publication bias.

One can only diagnose a publication bias in a discipline in connection with statistical power, including the probability to reject the null hypothesis correctly. According to Cohen (1962), the statistical power in psychological studies is actually around 60%. This means in the long-run, one can expect 60% significant results, given the stated hypothesis are all true, and all results would be reported. In reality, Fanelli showed that 92% of all published results reached significance (2010). This shows that the high number of positive results can possibly be explained by a publication bias. However, the fact that a majority of psychological research reports more positive findings, says nothing about whether these results are ‘true’ or ‘false’.

2.2. Estimating the reproducibility of psychological science

As mentioned in the introduction, in 2015, an international team of 270 researchers had the aim to estimate the reproducibility of psychological science (Open Science Collaboration, 2015). Thus, the scientists dealt with replications, which describes the attempt to recreate conditions to get a result that has already been shown. Within this thesis a replication will be defined as a tool to estimate reproducibility (Schmidt, 2009).

Over three years, scientist of the open science collaboration replicated 100 experimental and correlative studies out of three major psychological journals (Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP:LMC), Psychological Science (PSCI) and Journal of Personality and Social Psychology (JPSP)). The team selected high-powered studies, for which the material was accessible. Further, authors of the original work were contacted and interviewed. In the selected studies, the original results showed 97% significance; in the replications only 36% of the studies reported a *p*-value below the significance threshold of .05. In more than half of the studies the results could not be replicated (see Table 1).

Table 1. *Summarized replication results of the reproducibility project (Open Science Collaboration, 2015).*

	Journal	Replications p < .05 in original direction	Percentage	Average replication power
	Overall	35/97	36%	0.92
cognitive	JEP:LMC	13/27	48%	0.93
	PSCI	8/15	53%	0.94
social	JPSP	7/31	23%	0.91
	PSCI	7/24	29%	0.92

In addition, the authors observed that the reproducibility in cognitive psychology research performed better (13/27 and 8/15) than in social psychology research (7/31 and 7/24). This can have various reasons from the hypothesis to the type of test. For example, it was easier to replicate effects that tested main or simple effects (23/49) and particularly challenging to replicate interaction effects (8/37) (Open Science Collaboration, 2015, p. 5).

In order to verify the replication success, statistical characteristics were used, such as the *p*-value and the effect size. Figure 2 shows their distribution in violin plots. In comparison to box plots, violin plots include a core density estimation.

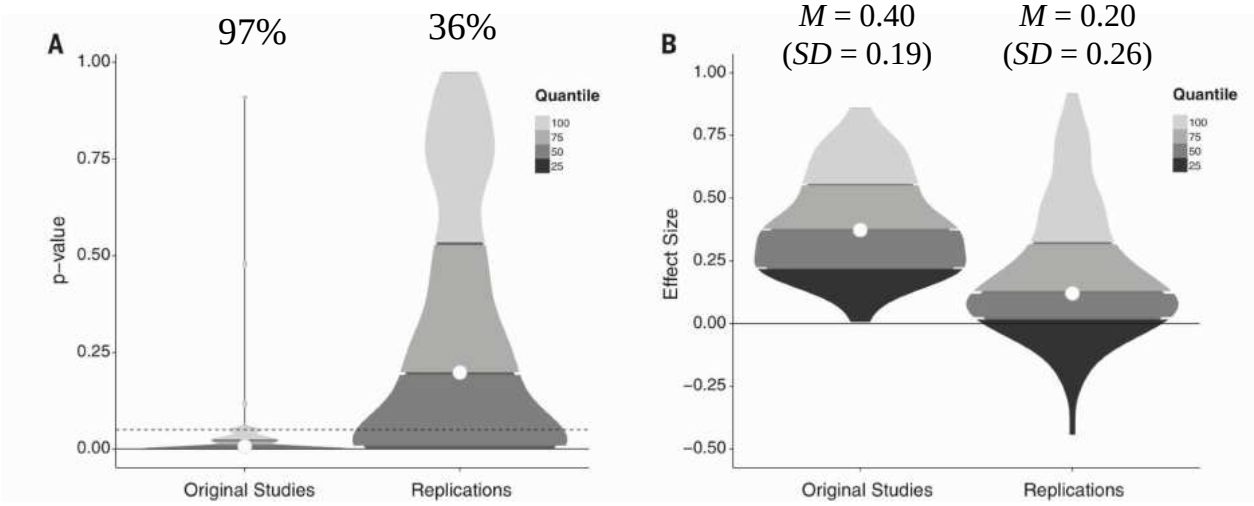

Figure 2. Violin plots of original and replication *p*-values and effect sizes (Open Science Collaboration, 2015). (A) *p*-values. (B) effect sizes. Lowest quantiles for *p*-values are not visible because they are clustered near zero.

Figure 2A shows that in the original studies nearly every *p*-value lies under the significance limit of .05 (97%) and in the replications only 36% of the *p*-values are significant. In Figure 2B the distribution of the effect size *r* is shown. The average

effect size of the replications (M = 0.20, SD = 0.26) is only half as large as the average effect size reported in the original studies (M = 0.40, SD = 0.19). In fact, it even shifts in the direction of effects that are observable. 82 of 100 effect sizes were larger in the original studies and both distributions correlated positively.

These indicators show the deficiency of replications, but it is important to note that no single indicator adequately describes the replication success. Another option, for example, is to look at the confidence intervals (Open Science Collaboration, 2015).

2.3. Questionable research practices (QRPs)

John, Loewenstein, and Prelec (2012) asked 2000 psychologists for their involvement in QRPs. The percentage of respondents to QRPs was very high, 94% agreed in at least one of the ten items (see Figure 3). The authors used the Bayesian-truth-serum (BTS), which provides incentives to tell the truth, and it led to the fact that in spite of their own involvement in QRPs, respondents likely told the truth. In comparison, the self-admission rate in the control group amounted to only 33% (John et al., 2012).

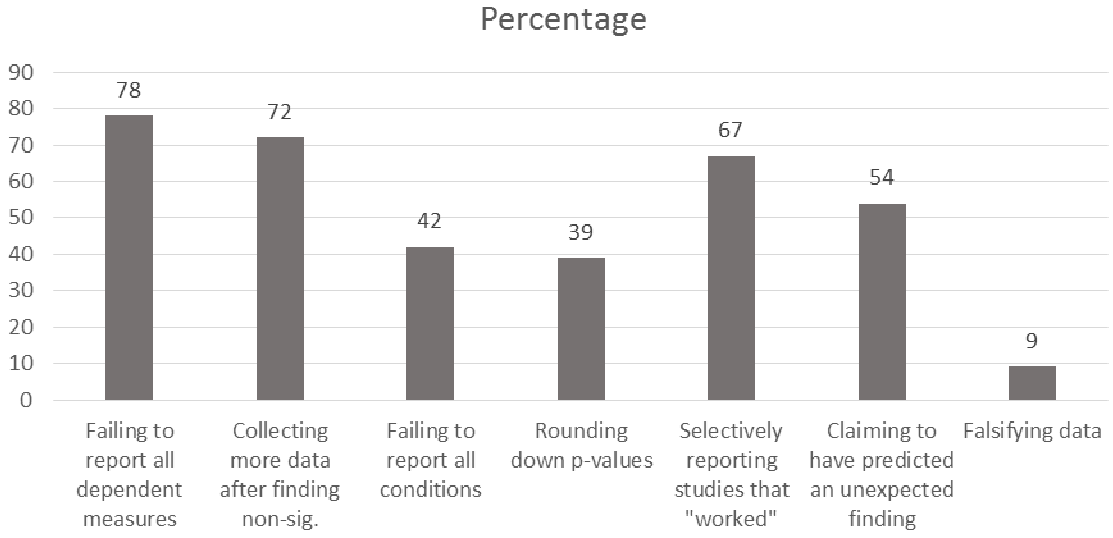

Figure 3. Percentage of questionable research practices (John et al., 2012). These results of the Bayesian-truth-serum condition in the main study show the geometric mean of three rates: self-admission rate, prevalence estimate rate and prevalence estimate rate derived from the admission estimate (i.e. self-admission/admission estimate).

Figure 3 shows, for example, that 39% of the respondents admit rounding down *p*-values and even 9% falsifying data. This influence on the analysis process and on the *p*-value, to reach significance, means that *p*-hacking took place. Even, if there are sometimes good reasons for violations, for example, to exclude certain data

(John et al., 2012), this study shows again publication pressure and that scientists want to publish remarkable results.

2.4. Prevalence of *p*-values under .05

Masicampo and Lalande (2012) investigated the psychological literature to consider whether the lived standard of the significance limit influences the *p*-value distribution. They analysed over 3627 *p*-values reported in three well-known journals (Journal of Experimental Psychology: General (JEPG), Journal of Personality and Social Psychology (JPSP), Psychological Science (PS)). Figure 4 shows their results. The *p*-values range from .01 to .10 on the x-axis and contain the critical value of .05, and show the frequencies of the *p*-values in intervals of .00125 on the y-axis. The authors state that the number of *p*-values, which just fall below .05 is higher than the predicted number of *p*-values based on the overall distribution of *p*. This points out *p*-hacking and an over-emphasis of the statistical significance (Masicampo & Lalande, 2012). The authors' inference is that researchers push their obtained empirical *p*-value just below the .05 limit. As a result, *p*-values accumulate just below the threshold of significance.

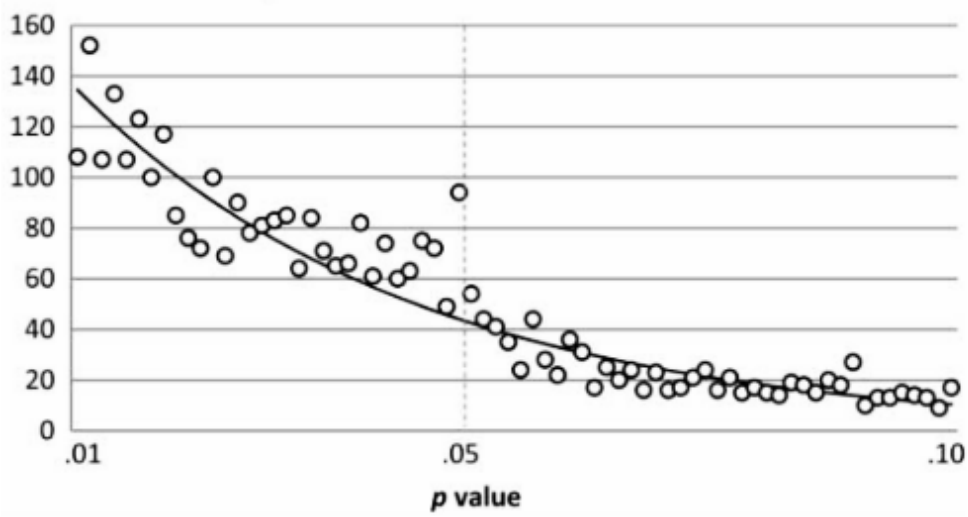

Figure 4. *P*-value frequencies at divisions of .00125 (Masicampo & Lalande, 2012).

At first, the use of these results to identify *p*-hacking seems to make sense. However, on closer investigation the problem is more complex than it seems. In fact, Masicampo's and Lalande's results are challenged by Lakens (2014b, p. 2). According to Lakens, four vital factors determine the *p*-value distribution. First, the number of studies examining true effects and false effects; second, the power of

studies that examine true effects; third, the frequency of Type I error rates and; finally, publication bias. These factors, he argues, would need more data collection for a more accurate interpretation of the *p*-value distribution (Lakens, 2014b, p. 5). Indeed, Lakens claims:

> 'P-hacking does not create a peak in p-values just below .05. Actually, p-hacking does not even have to lead to a left-skewed p-curve. If you perform multiple independent tests in a study where the null-hypothesis is true the Type one error rate is substantially increased, but the p-curve is uniform, as if you had performed 5 independent studies.' (p. 3)

Further, he asserts that a publication bias should result in significantly less *p*-values appearing above the significance limit, that true effects are causing a right-skewness and that an inflated Type I error rate produces a left-skewness for *p*-values below .05 (Lakens, 2014b, p. 2).

The research of Lakens suggests that the sole consideration of the *p*-value is not sufficient to make statements about data distortions and to interpret them. In fact, the statistical characteristics and conditions behind should be coded too, like sample size, respectively the statistical power, the effect-size and Type I and Type II error.

Due to this indentified coding error of Masicampo and Lalande (2012), a more detailed look at the *p*-curve including vital factors introduced by Lakens should be considered. Simonsohn et al. (2014) present the *p*-curve as a modern diagnostic instrument to discuss publication bias.

2.5. *P*-curve as a modern diagnostic instrument

The *p*-curve method by Simonsohn et al. (2014) takes on the *p*-value distribution of several significant *p*-values ($p < .05$), which is proposed as the key to the file drawer problem. Here, the authors assume that if there were no true effects examined in a series of studies, all *p*-values would be uniformly. In the case of the range of *p*-values between 0 and .05, this would mean that the frequency of *p*-values for each of these intervals is 20%. All *p*-values would occur equally frequent, like the red dotted line 'null of zero effect' in Figure 5 indicates. Further, they suggest that for examined effects the *p*-curve should be right-skewed. The *p*-curve should contain more low (.01s) than high (.04s) significant *p*-values, since researchers should strive to achieve highly significant results instead of just significant results (Masicampo & Lalande, 2012). If researchers just aim to achieve significance, the *p*-curve would be

left-skewed. There would be more high (.04s) than low (.01s) significant *p*-values, due to applications of QRPs, such as stopping rule. Left-skewness would, for instance, indicate a stop in data collection as soon as it is significant and would show limited ambition of data-minded research. Thus, only right-skewed *p*-curves are diagnostically strong: the more statistical power, the more right-skewed. The rule of thumb says that statistical power is good, when the curve is led with 80% of low *p*-values. Finally, the pattern of *p*-curve reflects the ambition of researchers.

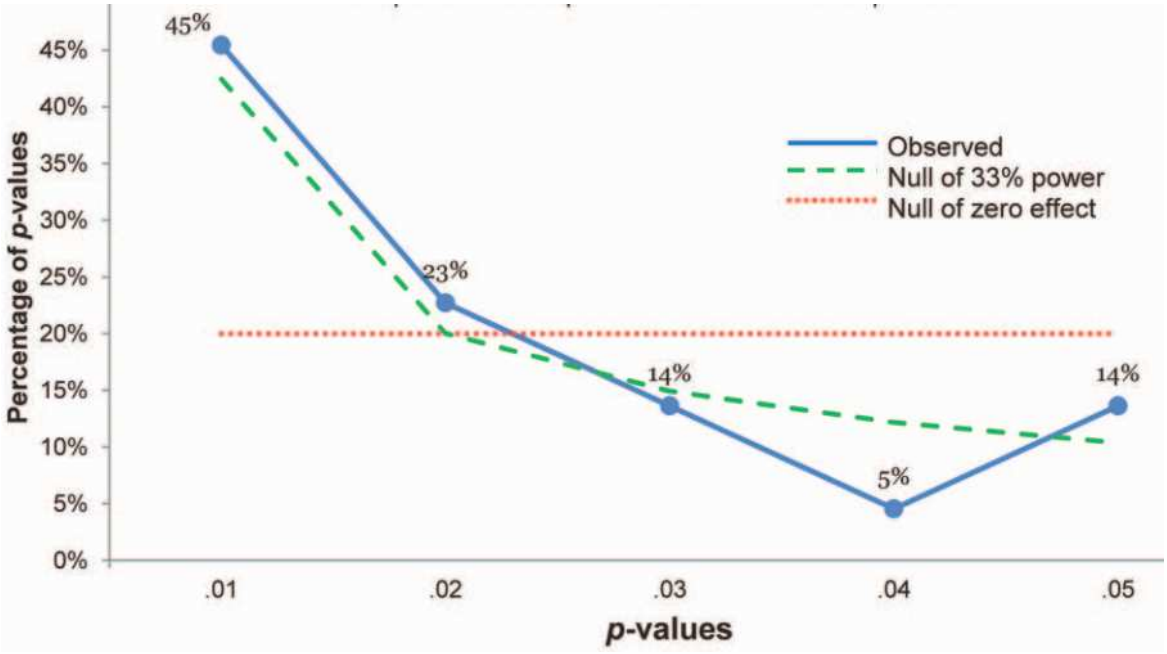

Figure 5. *P*-curve with a sample size of 20 *p*-values out of the Journal of Personality and Social Psychology (Simonsohn et al., 2014).

The blue lined *p*-curve in Figure 5 shows low statistical power because the curve is led with 45%. Also, it shows an increase of *p*-values at .05. These two points indicate a possible *p*-hacking (Lakens, 2014b; Masicampo & Lalande, 2012).

Furthermore, Simonsohn et al. (2014) point out that some points must be considered with regard to *p*-values. The curve or values, for example, only relate to specific statistical analyses and not to theory. Results can be significant, while the theory is wrong and results can be non-significant while the theory is true. Series of studies may be incorrect due to specific operationalization and specific samples. They state "[…] selected *p*-values must be (a) associated with the hypothesis of interest, (b) statistically independent from other selected *p*-values, and (c) distributed uniformly under the null" (Simonsohn et al., 2014, p. 535). Finally, the authors propose to report a *p*-curve disclosure table, which is also recommended by Lakens (2014b, p. 4). The p-curve disclosure table is intended for improving the reliability and

reproducibility of *p*-curve analyses. According to Simonsohn et al. (2014) the disclosure table is defined by five aspects which, in turn, lead to the selection of *p*-values and an accurate creation of a *p*-curve:

1. Identify the researchers' stated hypothesis and study design
2. Identify the statistical result testing the stated hypothesis
3. Report the statistical results of interest
4. Recompute the precise *p* value(s) based on reported test statistics
5. Report robustness results (p. 541)

If compared to the *p*-value distribution of Masicampo and Lalande (2012), it is obvious that this contemplation takes more aspects into account. Hence, this master thesis will include parts of these within the methods section.

2.6. Statistical Power analysis

Kühberger, Fritz, and Scherndl (2014) presented another concept for identifying data distortion in research processes. They draw the attention to statistical power analysis, which calculates the appropriate sample size for the hypothesis. In order to calculate the appropriate sample size, the desired effect size and the selected analysis is considered. Scientists should strive for possessing the appropriate sample size because it increases statistical power and thus the probability that a test correctly rejects the null hypothesis. The larger the sample size, the more accurate the distribution. Hence, to reduce the Type II error the sample size must be increased. According to Simmons (2013, p. 1) the criticism, that a trivial effect gets significant with a large sample size, makes no sense, since "the effect must be trivial and we don't care about trivial effect sizes". Accordingly, Kühberger et al. (2014) claimed that if researchers conducted a statistical power analysis before data collection, then effect size and sample size would correlate negatively. This is due to the fact that for an expected big effect of an investigation, one needs a small sample size and for an expected small effect, one needs to find a large sample size.

Within their research Kühberger et al. (2014) asked 1000 psychologists for their examinations and whether they have or have not performed a statistical power analysis. Surprisingly, the answer was negative, statistical power analyses were not applied in most cases. Consequently, sample size and effect size should be independent of each other, but the negative correlation nevertheless emerged in this case with r = -0.45, 95% *CI*: [-0.53; -0.35] (see Figure 6).

Due to any application of statistical power analysis, this may not be the explanation. However, as mentioned in the part of the *p*-curve, scientists can apply the stopping rule. They can “collect more data after finding non-significance”, which corresponds to an item of QRPs (John, Loewenstein, & Prelec, 2012), and then they can stop as soon as the result is significant. Further, they can exclude or not report certain cases to achieve significance (Kühberger et al., 2014). These facts would refer to *p*-hacking. Based on the publication bias only studies make it to the literature which are really significant or have somehow gained. Yet, one could also report an unexpected finding retrospectively as a hypothesis. Hence, *p*-hacking and publication bias come into question as a cause for the negative correlation. This result shows that researchers rather dedicate minimal effort to achieve significance, regardless of their findings true strength which would be expressed with the effect size. Indeed, the relation that with smaller effects larger samples sizes are required is often forgotten.

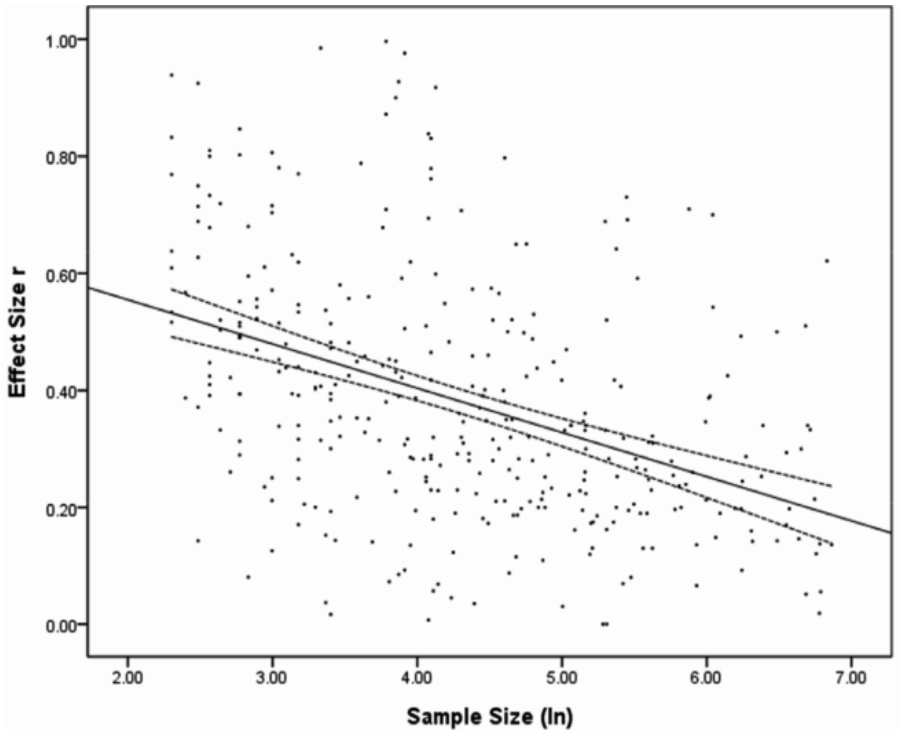

Figure 6. Corrected effect size r plotted against logarithmically transformed sample size. Negative correlation of *r* = -.45, 95% *CI*: [-0.53; -0.35], between sample size and effect size (Kühberger et al., 2014).

In short, the correlation of sample size and effect size of Kühberger et al. (2014) shows another possibility to verify data distortion. For the present work, this shall demonstrate as well as highlight the importance of a pluarist approach. As many

indicators as possible will be considered for the applied analyses to get comprehensive support for conclusions in terms of data distortion.

3. Research questions

The aim of this master thesis is to test the described methods which were used for statistical investigations to point out QRPs. As mentioned in the introduction, the sample for the analyses consists of student theses in the field of economic psychology. Possible reasons can be found for and against the findings of data distortion in students work through QRPs. This research assumes that there is less data distortion to be found, since students are not exposed to publication pressure like scientists are. On the other hand, if the findings hint towards QRPs, causes for these practices might lie in the teaching process rather than publication pressure. Hence, the research question is:

Is there evidence for questionable research practices amongst students' theses?

After having discussed the theory it is evident that further questions are necessary to answer the stated research question. Hence, this research aims to debate five concerns that emerged during the discussion of the theory. Other inquiries are as followed:

1. Does the *p*-value distribution point towards *p*-hacking?
   Note. As a comparison to the *p*-curve of Simonsohn et al. (2014).

2. What is the average power of the studies? Is the publication bias observable?
   Note. As a comparison to the p-curve of Simonsohn et al. (2014) and the calculated power of Cohen (1962).

3. What does the distribution of p-values and effect sizes look like?
   Note. As a comparison to the violin plots of the Open Science Collaboration (2015).

4. Is there a correlation between sample size and effect size?

   Note. As a comparison to the conducted correlation of Kühberger et al. (2014).

5. Which errors are observable in the reporting?

The next section considers the exact sample size, calculation methods and the coding scheme as the most appropriate method to answer these questions. Specifically, question 1 and 2 will be examined with regard to the *p*-curve, question 3 asks for violin plots, question 4 for a correlation and finally, question 5 will be explored descriptively.

4. Method

At first, it was necessary to identify theses in the fields of business, organisational and economic psychology. 692 student theses have been supervised officially since 1993. 461 of these were attributed to the field of economic psychology as defined by a supervision or co-supervision by Prof. Dr. Erich Kirchler. With the comparison of supervisor lists and through a re-examination, the original list was revised and supplemented with missing content and finally amounted to 513 theses from 2016 to 1993 in the field of economic psychology. 151 of these theses were available in digital form in the archive of the Department of Applied Psychology: Work, Education and Economy. 20 further theses were found online as PDF-files with E-Theses of the University of Vienna. Six additional theses have been received via Facebook and email, so overall, there were a total of 177 theses available in digital form. The availability of the remaining theses was assessed in the departmental library and reviewed as printed originals.

Since some work was written in collaboration, used a similar set of data, had similar hypotheses or were not available in the library, some had to be excluded from the list. Altogether, there was the possibility and availability of a total of 431 student theses to be coded from 2016 to 1993 (see Table 2).

Table 2. *Number of theses per year, which were available (online or in departmental library).*

year	number	year	number	year	number
1993	2	2001	19	2009	28
1994	6	2002	22	2010	17
1995	20	2003	29	2011	19
1996	19	2004	13	2012	18
1997	26	2005	19	2013	13
1998	16	2006	18	2014	12
1999	13	2007	24	2015	26
2000	16	2008	27	2016	9

To limit the coding of the available theses to a feasible amount of work, all available theses from 2000 to 2016 were coded for the present study, which amounts to a total of 329 theses. 67 of these theses were verified by one rater and the rest of 262 theses were coded by two raters to increase reliability.

4.1. Sample size

In order to collect the different characteristics and statistical data of student's theses, SPSS was used for developing a coding scheme. The first step of the coding scheme was to classify the type of theses (in comparison, see *Decision tree for classification* of Kühberger et al., 2014, p.3). It was necessary to identify whether the thesis is a non-empirical, a qualitative, a descriptive or an explorative one. After doing so it could be determined if it should be excluded and to identify hypothesis testing studies to include them (see Table 3). A total of 259 theses could be included, 70 had to be excluded.

Table 3. *Research method classification and number of theses.*

N	Classification
5	non-empirical
34	qualitative
4	descriptive
27	exploratory
259	included

A second important feature was the coding of the analysis to test the hypothesis. On the one hand because the analyses have different corresponding statistical characteristics, like test values, which need to be coded, and on the other hand to calculate and check the *p*-values and effect sizes afterwards with the right technique (see Table 4).

Table 4. *Numbers of analyses of all included and usable theses inspired by Kühberger et al. (2014).*

N	Analysis
23	Tests for categorical data (chi square)
11	Tests for means (t-test)
132	Tests for means (F-test)
20	Correlations
46	Linear regression
11	Tests for medians (non-parametric)
11	Complex analyses
254	total

Note. Due to faulty result presentations and incomplete data five more theses had to be excluded. Therefore, the total sample size of this master thesis consists of 254 student theses in economic psychology, which however varies within the individual analyses.

4.2. Calculations

Before *p*-values and effect sizes had been involved in the analyses, they were checked for errors. For their recalculation, it was necessary to do a statistical research to find appropriate calculators and techniques. These allow their recalculation with certain characteristic values and without having access to the whole data set of the corresponding hypothesis.

4.2.1. Calculation of p-value

P-values were recalculated from various test scores with the ‘Quick *P*-Value Calculators’ of the homepage ‘Social Science Statistics’ (Stangroom, 1999), which provides statistical resources for students. Specifically, for *p*-values of variance tests and *t*-tests an excel sheet of Lakens (2013) was used. As an example, the *F*-score and the degrees of freedom (*df1, df2*) were entered and the corresponding *p*-value and the eta square as the effect size were automatically calculated.

4.2.2. Calculation of effect size

Effect sizes were recalculated with the homepage of Lenhard and Lenhard (2016), where all effect size calculations are backed up with scientific literature (Fritz, Morris, & Richler, 2012, p.12; Elis, 2010; Borenstein, Hedges, Higgins, & Rothstein, 2009; Cohen, 2008; Morris & DeShon, 2002, p.119; Rosenthal & DiMatteo, 2001; Rosenthal, 1994, p. 239; Cohen, 1988, cited by Lenhard & Lenhard, 2016). Effect sizes were converted into *r* because its efficiency within the field, its limited range of -1 to +1 and its more accessible interpretation (Open Science Collaboration, 2015, p. 3). Specifically for effect sizes of variance tests and *t*-tests the excel sheet of Lakens (2013) was used again.

4.2.3. Calculation of statistical power and p-curve

To calculate the statistical power of the studies and to create a *p*-curve the *shiny application* of RStudio (2013) was used. The *shiny application* of RStudio is a web application framework for R. Finally, having applied the calculators and the excel sheet of Lakens (2013) it was possible to determine which statistical characteristics were needed and finally, the coding scheme could be completed.

4.3. Coding scheme

For the identification and coding of the right study classification, the right hypothesis and analysis, the corresponding *p*-value, test value, and so on, as mentioned in part 4.1 the papers of Kühberger et al. (2014, p.3) and Simonsohn et al. (2014, p. 340-342) were used. In total 74 characteristics were ultimately assessed in a single thesis and transferred into the created SPSS coding scheme.

With regard to the selected hypothesis out of the student thesis, the salient hypothesis was chosen or alternatively the first, which was reported. In addition, this research assessed whether a statistical power analysis was applied and whether the *p*-value was reported in three or two digits or in a range. Due to length, in Table 5 some of the characteristics are listed with their purpose of usage (for the whole coding scheme see appendix).

Table 5. *Characteristics and purposes for the coding scheme.*

Characteristic	Purpose of usage
Year	To examine changes over the years
Classification	To include or exclude study
Hypothesis	Salient or first one to select right *p*-value
Sample size criterion	Power analysis (yes, no or rule of thumb)
Analysis	For selecting the right statistical values
p-value	For *p*-curve analysis and distribution
Effect size	For checking distribution and correlation
All different test values (*F, t, r, β, X²,…)*	To recalculate *p*-values and effect sizes without whole data set
Reporting	Is something missing?
Exploration	Was there an exploration section?

5. Results

The structure of the result section is as follows. First, the descriptive results of *p*-values and effect sizes are presented and their deviations after recalculation. Second, to answer the first research question 'Does the *p*-value distribution point towards *p*-hacking?' and the second 'What is the average power of the studies? Is the publication bias observable?' the *p*-curve is explored. Next, with regards to the third question, both statistical characteristics are compared in violin plots with the results of the reproducibility project (Open Science Collaboration, 2015). Afterwards, in order to assess the fourth research question, the correlation between the sample size and the effect sizes is investigated. Finally, the errors in statistical reporting are analysed exploratively with regard to the last research question.

5.1. Descriptive results

5.1.1. P-value distribution

The descriptive results of the *p*-values and their distribution are illustrated below in Table 6. The first column shows that in total 244 *p*-values could be found in the sample of 254 studies and that they were not reported 10 times. Also, the table shows that 178 of these *p*-values were reported exactly to three or two digits and 66

were reported in a range. With regard to all 254 cases, it was possible to recalculate 183 *p*-values. It shows that there were cases, where the original *p*-value was reported to a range or not at all. After recalculation, the same *p*-value was available exactly to three digits. Thus, the maximized column consists of all recalculated *p*-values and also *p*-values, for which the recalculation was not possible.

Table 6. *P-value distribution and deviations.*

	p-value	checked	maximized	deviations
3, 2 digits	178		231	23
range <.05, <.01	64	183		3
range >.05	2			
Σ / %	*244 / 96%*	*183 / 75%*	*231 / 91%*	*26 / 14%*
Missing's	10	71	23	13 with .01
Total	254	254	254	

Deviations in the 183 recalculated *p*-values were identified in 26 cases (14%) but half of them (13/26) had only a deviation of .01. The corrected *p*-values were included. However, the 66 *p*-values reported to ranges were excluded, due to possible frequency distortion they would cause in the *p*-curve analysis.

With regard to the original *p*-value variable 142 of 244 *p*-values (58%) are significant. In the maximized variable 121 of 231 *p*-values (52%) are significant. This looks as if after the *p*-value assessment less *p*-values are significant. However, this is not the case since the original *p*-value variable contains *p*-values, which were specified as significant and reported in a range (<.05, <.01) and could not be recalculated exactly. The sample size is important, because the *p*-curve is only considering significant *p*-values. Hence, the sample size for the *p*-curve consists of 121 significant *p*-values, which could get assessed or were reported accurately.

5.1.2. Effect size distribution

The effect size distribution and their deviations are seen in Table 7. The first column shows which effect sizes were reported and how many times. In sum 98 effect sizes were found in student theses (39%) and 156 were missing. The second column shows the total of the assessed effect sizes and which effect sizes could be subject to recalculation. The maximized column shows which effect sizes could be converted into *r*. Thus, the final sample size consists of 228 effect sizes (90%).

Seven deviations were found regarding to the original effect size variable (98 reported effect sizes) and again half of them (4/7) had only a deviation of .01.

Table 7. *Effect size distribution and deviations.*

	effect size	checked	maximized r	deviations
r / d / n² / r²	21 / 2 / 29 / 6	0 / 31 / 129 / 1	32 /31 / 129 / 1	0 / 1 / 5
Stand. beta	32	0	32	1
Unstand. beta	4	0	0	0
Cramer's V	1	1	1	
Odds Ratio	3	0	2	
Σ / %	*98 / 39%*	*162 / 64%*	*228 / 90%*	*7 / 7%*
Missing's	156	92	26	4 with .01
Total	254	254	254	

5.2. *P*-curve

By running a *p*-curve analysis, the first and second research questions will be answered. Figure 7 shows the results. The *p*-values range from 0.0 to .05 on the x-axis and show the percentage of *p*-values on the y-axis.

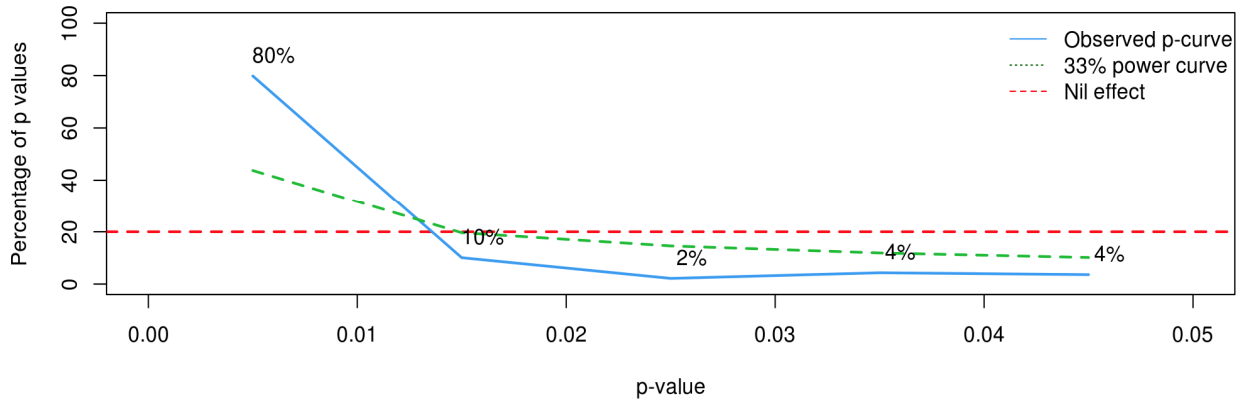

Figure 7. *P*-curve with *N* = 121 significant *p*-values of the maximized variable and a median observed power of 64.81%.

In fact, studies contain evidential value because of its right skewness (z = -28.8, p < .001). Since the curve is led with 80% of low *p*-values (z = 22.3, p = 1) the evidential value here is not insufficient and thus, the studies do not lack of evidential value and did not get *p*-hacked because the curve is not left skewed (z = 28.8, p = 1). Hence, the first research question 'Does the *p*-value distribution point out to *p*-hacking?' can be answered with no, because the curve is right skewed and shows no

further increase before .05. It seems that the significance limit of .05 does not influence the students to *p*-hack their data.

Part one of the second question ‘What is the average statistical power of the studies?’ can be answered with *Mdn* = 64.81% (median observed statistical power). This post-hoc statistical power analysis shows that the null hypotheses were correctly rejected with a probability of rounded 65%. That power is similar to the calculated power of Cohen (1962), which was 60%, and it means that 65% significant results can be expected to be found in the studies of the student theses, considering that the stated hypotheses were all true. The low statistical power points towards too small sample sizes, since only appropriately large sample sizes can increase the power and detect effects or can reduce Type I and Type II errors. As mentioned in the descriptive result section ‘*p*-value distribution’, there were not 65% significant *p*-values but 52% significant *p*-values for the hypotheses. Hence, it also can be assumed that the proportion of non-significant results increases if examined hypotheses are false.

Moreover, the findings respond to the second part of the question ‘Is the publication bias observable?’ with no. The number of significant results even falls below the statistical power. Students just reported their results, no matter whether they were significant or not. In conclusion of this section it can be argued that the publication bias, which describes the preferred publication of studies with ‘positive’ or significant results, did not occur.

5.3. Violin plots

The third research question asked ‘What does the distribution of *p*-values and effect sizes look like?’ and will be discussed in Figure 8 and 9. In the two figures, the violin plots of the *p*-values and effect sizes are compared with the results of the reproducibility project (Open Science Collaboration, 2015).

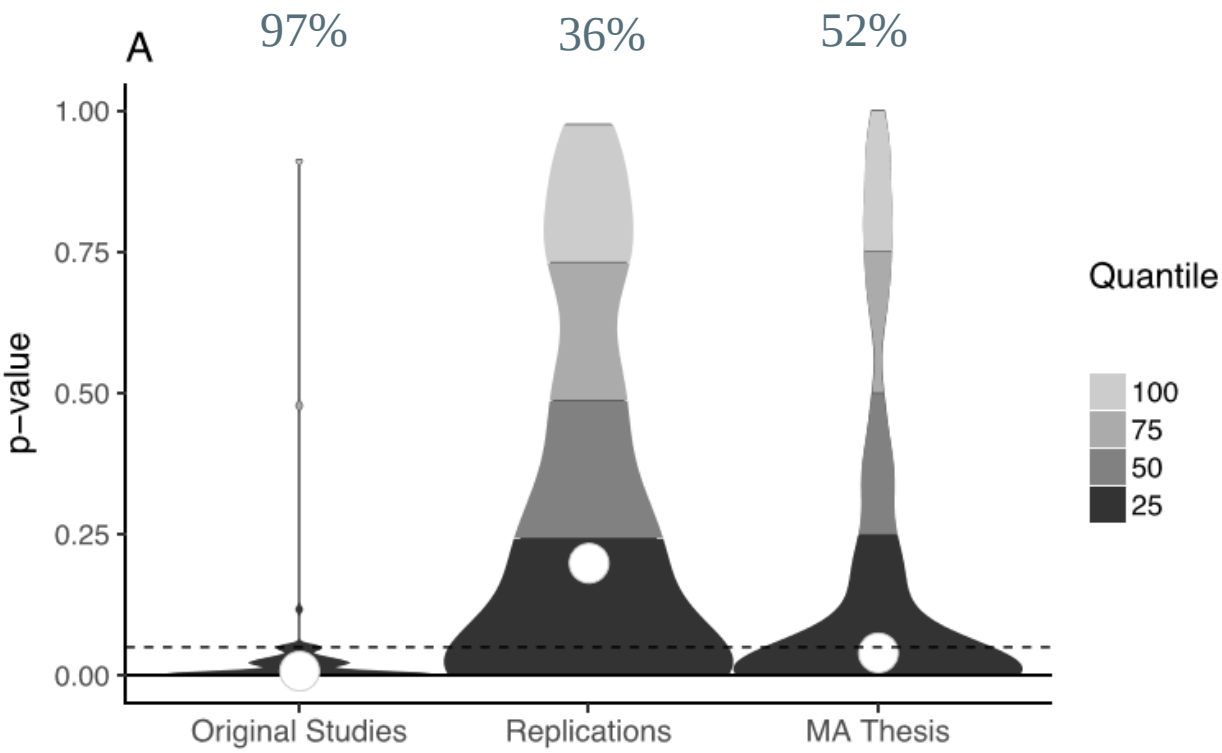

Figure 8. Violin plots of original, replication and master thesis *p*-values (A). The density shows the number of *p*-values. Lowest quantiles for *p*-values are not visible because they are clustered near zero (Open Science Collaboration, 2015).

The first two violin plots in Figure 8 of the *p*-values were presented before in the theoretical background section in Figure 2A, the plot on the right adds the *p*-values of the student theses, for a direct comparison. As shown the students' work are distributed similar to the *p*-values of the replications. Due to statistical power 65% significant results were expected, however less than that, namely 52% (121 of 231) significant results were found. Based on this data, it can be concluded that no selective reporting took place. Further, the significant results found mirror the results of the replications.

Figure 9 shows the violin plots of the effect sizes. On the left are the two plots presented before in Figure 2B and on the right the effect sizes of the student theses are shown. The form of the 'MA Theses' resembles again the form of the replications. Further, the mean value with M = .23 (SD = 0.18) is similar to that of the replications with M = .20 (SD = 0.26). The two obervations support once more that no selective reporting took place. For better comparability, only the mean value is reported, and not the median, since the Open Science Collaboration also only reported the mean value.

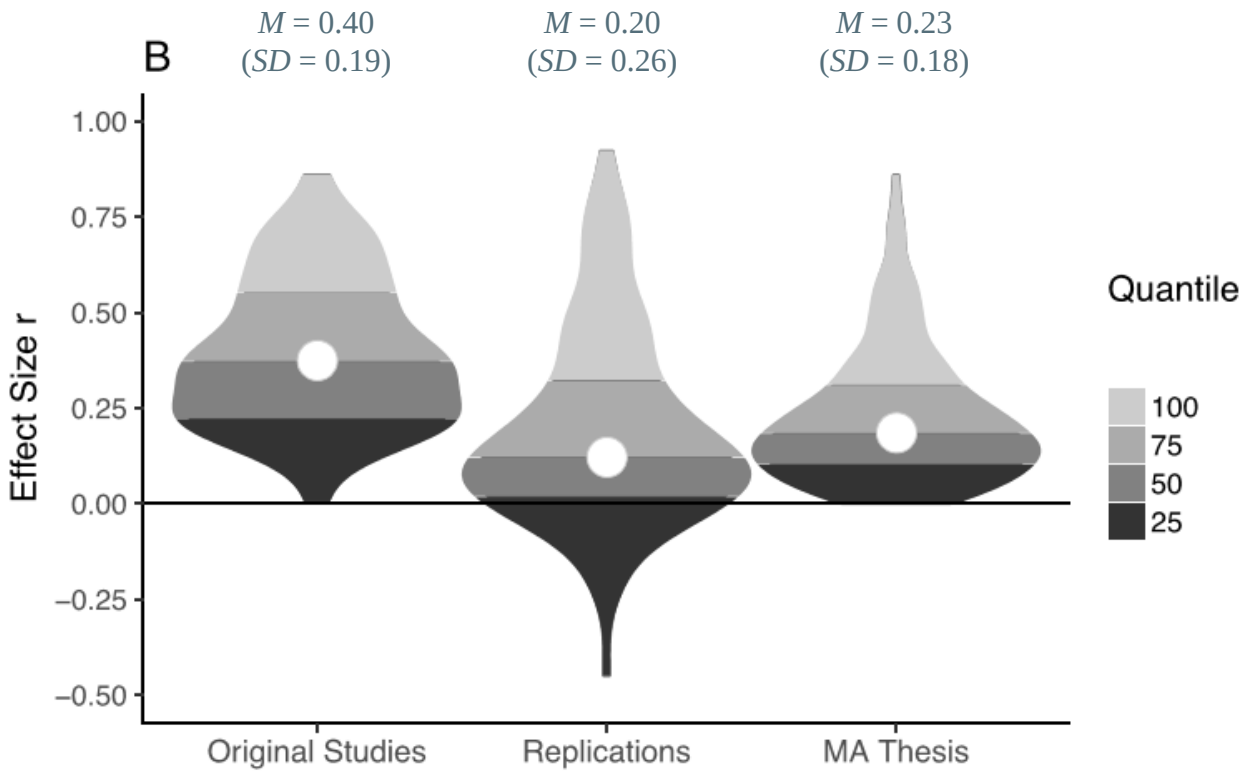

Figure 9. Violin plots of original, replication and master thesis effect sizes (B) (Open Science Collaboration, 2015).

5.4. Correlation

To answer the fourth research question 'Is there a correlation between sample size and effect size?' a Pearson-correlation was computed. According to Kühberger et al. (2014) a negative correlation should be found when scientists apply statistical power analyses. However, statistical power analyses did not take place and hence the correlation indicates data distortion. Among the students only one person did conduct a statistical power analysis. In this case, if no data distortion took place, effect size and sample size should be independent from each other. Figure 10 shows the result. Between the effect size and sample size of the student theses no correlation could be found, r = -0.07, 95% *CI*: [-0.20; 0.06], p = .282, which fits the observation. The independence of these two sizes supports Kühberger et al. (2014) observations and concludes that no distortion took place.

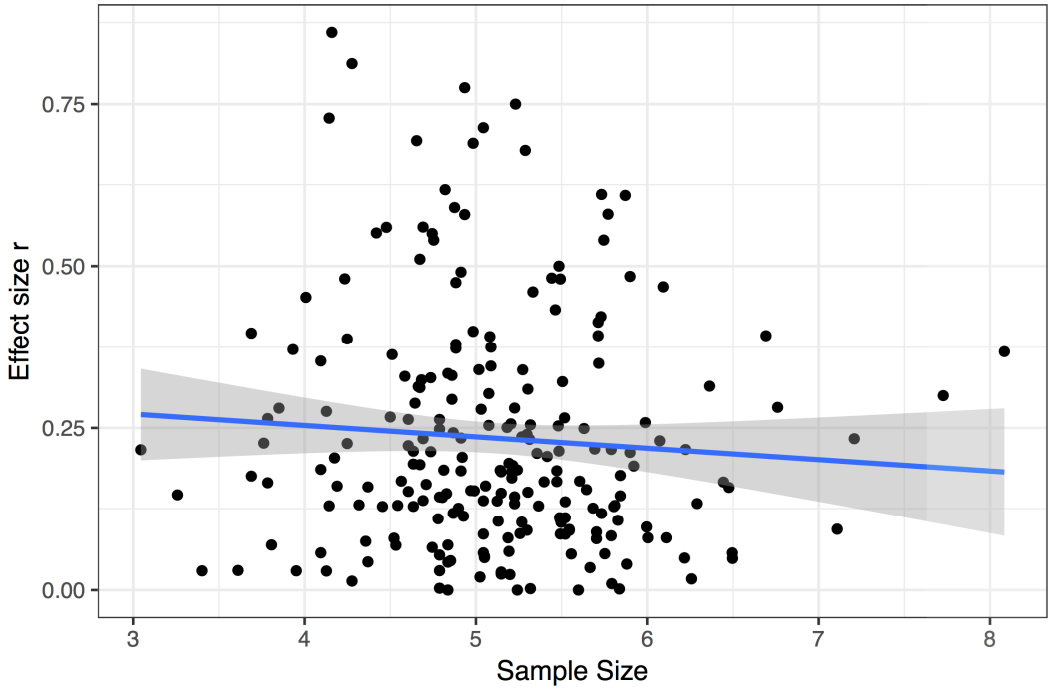

Figure 10. Pearson-correlation of effect size and logarithmized sample size of the student theses with r = -0.07, 95% *CI*: [-0.20; 0.06], p = .282.

5.5. Reporting errors

'Which errors are observable in the reporting?' is the last question to be answered. The findings show that errors occurred in half of the theses, namely in 136 of 254 cases (54%). The error rate does not necessarily include the not reporting of an effect size. This was not considered as an error in the first years. However, the error rate includes, for example, when a student forgot to list the table of content in his thesis or when his hypotheses were flawed. Consequently, also mistakes that were not related to the statistics were coded as reporting errors. Some observed and summarised results are shown in Table 10, which presents absolute numbers of missings' in the theses. It should be noted that several errors could also occur together in one thesis.

Table 8. *Reporting errors in absolute numbers.*

Statistical characteristics	Missing's
p-value	10
Effect size	156
Test value	13
df	29
N	0
M (SD) of age *N*	36 (50)

5.6. Exploration

The exploration has shown further interesting results. Something have changed in student theses over the last 16 years. Table 8 shows a few correlation results. One can see that the total page number has decreased with r = -0.45 (p < .01), which is almost a large effect. Also, the page number of the result section has decreased with r = -0.34 (p < .01), which corresponds to a moderate effect. As seen in the negative correlation of r = -0.20 (p < .01), the reporting of an effect size has increased over the years. It should be stated here that coding '1' is representative for an effect size being reported and '2' that no effect size was reported. Only for the sample size, there is no significant correlation.

Table 9. *Exploratory correlation results.*

	Correlation *r* with years
Sample size *N*	0.023
Total thesis page number	-0.451**
Page number of the result part	-0.341**
Effect size	-0.199**

Note. *N* = 254.
** *p* < .01.

To further illustrate these correlations, Table 9 shows the mean values and standard deviations of the first and the last two years, including the four aspects out of Table 8 'Exploratory correlation results'. Although, on first observation it seems as if the sample size N has decreased, it has not. This is due to an outlier in the first two years and can be observed in the standard deviation. The total number of pages and the page number of the result part have almost halved, which is mirrored in the means. Further, 64% of the students report an effect size nowadays, which leaves room for improvement.

Table 10. *First and last two years in comparison.*

	M (SD)	
Years	2000 – 2001	2015 – 2016
Sample size *N*	289.3 (648.1)	204.0 (166.9)
Total thesis page number	105.8 (47.1)	59.6 (21.7)
Page number of the result part	18.8 (7.5)	10.1 (6.7)
Effect size	22%	64%

Note. 2000-2001 *N* = 23, 2015-2016 *N* = 28.

5.7. Further testing

To demonstrate *p*-hacking, there are further methods applied in the psychological literature, for instance, considerations towards possible statistical influences by covariates (Asendorpf et al., 2013; Simonsohn et al., 2014). In general, there are good reasons to include covariates in calculations. However, in certain cases, scientists consider the result with and without the covariate, for the purpose of achieving and reporting selectively only the significant result (Simonsohn et al., 2014). Based on this, this research includes further descriptive testing on the data of student theses.

Table 11 shows the results. In total 44 of the 254 theses contained covariates (17%). 20 of these theses with covariates were non-significant, which corresponds to 45%, which means that 55% of the results with covariates were significant. Therefore, the remaining 210 theses without covariates contained significant results in 97 cases (46%). The ratio of significance for theses with and without covariates is found exactly reversed. This finding of a 9% difference shows the tendency that studies with covariates show more significant results, including the limitation of the small sample size of 44. Perhaps students only showed the significant result with the covariate because without, it was not significant, or perhaps they had a good reason to control a covariate. To make a clear conclusion, more data would be needed, like the result of significance without the covariate and a supporting theory.

Table 11. *Covariate and significance connection*

Covariate	Significance	
	Yes	No
With, *N* = 44	24	20
	55%	45%
Without, *N* = 210	97	113
	46%	54%

Note. 'Yes' means $p < .05$ and 'No' means $p > .05$.

Another interesting exploration result is that the majority of the students, namely 61% (74/121) did not include an exploration part in their thesis, if the result was significant, see Table 12. If the result was non-significant, 50% (67/133), half of

the students included an exploration part. Hence, it seems as if students engaged in further explorations more often, if results were non-significant.

Table 12. *Exploration and significance in results.*

		Exploration		
		yes	no	Σ
Result	significant	47	74	121
	non-significant	67	66	133
	Σ	114	140	254

6. Limitations

In this section possible limitations that occurred during the research are listed. First, a methodological limitation is that interrater reliability was not calculated for multiple raters which could have increased the reliability. Second, with regard to the existing significance limit, the literature considers the limit of .05 as overestimated and is getting criticized. However, in the master thesis this boundary was used. In the study for estimating the reproducibility, for example, *p*-values were also treated as significant if they were slightly above the limit of .05 (Open Science Collaboration, 2015, p. 3). Whilst this difference affects the percentage of significant studies it affects the *p*-curve less since it would only expand. Ultimately, it is important to have in mind that the significance limit offers room for interpretations.

Further, no results, neither replications nor explorations, are protected from the confirmation bias. In addition, as mentioned already in the theoretical background section, a result can be significant, even though the use of theory is wrong and vice versa. This also applies to this master thesis.

Another limitation emerges with regard to the estimation of statistical power, as applied to the *p*-curve distributions. Simonsohn et al. (2014, p. 539) recommend creating a rule to select studies for the comparison of *p*-values. In the master thesis, it has been determined that theses are limited to the field of economic psychology. One reason is to minimize subjectivity. Another reason is to minimize heterogeneity since no satisfying technique has been found yet to calculate power for studies with huge heterogeneity (Lakens, 2014a).

Moreover, with respect to the covariate testing, it would be interesting if the result is non-significant without the controlling for the covariate. However, it would be necessary to have access to the whole data.

Additionally, further distinctions are possible with regard to the identification of *p*-values. For example, a lot of students applied a manipulation check before testing their hypotheses. The influence of whether a manipulation succeeded or not on the significance of the main hypothesis could be raised. In this master thesis, this aspect was ignored.

Finally, there is the limitation of underpowered results, under which the whole field of psychology is suffering from. A larger sample size, in this case of student theses, is always desirable.

7. Statistical Guide

There are a number of existing recommendations for good research practices, as well as for open access databases and search engines (PsyDok, PsychData, Potsdam Mind Research Repository, PsychAuthors, Open Science Framework, ZPID, PsychSpider, PubPsych, PsychOpen, and so on). In addition, there are journals that even support the publishing of replications, like the 'Journal of Experimental Social Psychology' (Elisabeth, Veer, & Giner-Sorolla, 2016). Hence, it is obvious, that the relevance of the discussed topic is recognized.

For example, Fuchs, Jenny, and Fiedler (2012) suggest six requirements for scientists and four guidelines for reviewers and then asked 1292 psychologists from different countries whether they support them or if not, what their reasons are. The results show that in general the majority is 'open for change' (Fuchs et al., 2012, p. 639). However, psychologists tend to favour good practice standards instead of publication conditions. The only condition that is agreed by all is that authors should report all experimental conditions including those that have failed. Further, since most of the time rejection is justified by the fact that the requirement is not appropriate for all types of studies, the conclusion is drawn that customized standards for specific research types are needed (Fuchs et al., 2012, p. 640).

For this reason, the presented insight of this master thesis was used to create a customised statistical guide for students, see Table 12. The statistical guide is

based on the coding scheme and created with the help of students' emerging difficulties and challenges.

Table 13. *Statistical Guide for experimental and hypothesis testing studies.*

o Distinguish and clearly highlight research questions of hypotheses o Label correctly null hypothesis (H0) and alternative hypothesis (H1) o Define independent (IV) and dependent variable (DV)	
o Determine a priori the statistical characteristics: Error probability α = .05, β = .20, power 1-β = .80, select effect size o Determine appropriate analyses/test procedures for evaluation in order to perform a power analysis: www.gpower.hhu.de	
o Present the exact composition of the sample (why exclusion took place, …) o Sample description: *M* (*SD*) of age, *N* of cells, groups, ... (report all values)	
o Report all *p*-values exactly to 3 digits, also non-significant *p*-values	
o Report effect size (*r,* Cohens *d*, Glass' Δ, Hedges G, Cohens f^2, η^2, r^2, standardized *β*, unstandardized *β*, Cramers V, Odds Ratio, ...)	
o Report results with and without covariate/s and justify controlling for covariates o Report the test value for the whole model as well as for the predictors	
Analysis/test procedure	**Reporting**
Tests for categorical data: χ^2 -test (Chi)	χ^2-test value, degrees of freedom (*df*)
Tests for means: *t*-tests	*t*-test value, *df*, dep. or indep., report *t* correlation for dependent *t*-tests
Tests for means: *F*-tests (variance analyses, ANOVA, MANOVA)	*F*-test value, *df*1, *df*2, k = number of groups to be compared
Regression (lineare, multiple, mediations)	unstandardized *β*, standardized *β* (=effect size), report variance of IV and DV, if only non-standardized coefficients are reported

Tests for medians: non parametric procedures (Kruskal-Wallis-test, U-test, Wilcoxon-Mann-Whitney-test, ...)	selected test, test value, *df*, *z*-value
Correlation	*r*-test value
o Report *N, M, SD* and factorial cell mean values of the groups to be compared (for presentation create bar charts with confidence intervals, no SPSS outputs) o Caution: for non-parametric procedures report median and quartiles (for presentation create box-plots/violin-plots with quartiles, no SPSS outputs)	
o If the data has already been used, report it at the beginning	
o Report inter correlation of the IV and DV with *p*-value	
o Conduct further tests and/or explorations at the end and do not form hypotheses afterwards	

8. Discussion

To conclude, the assumption has been confirmed that there are less questionable research practices in student theses. The results of this master thesis give evidence for students' good research practices. However, as has often been mentioned the *p*-curve or statistical values refer to specific statistical analysis and not to the theory. It can be criticized that the samples are too small and that the reporting of the effect size is expandable.

Obviously, the topic of 'good research practices' is very comprehensive. It is not just about the scientists or students, it is also about the reviewers, editors and professors, who must ensure that honest research will be rewarded. The scientific culture should be aware of the pressure they generate with publishing only significant and individual results. At this point, further research could investigate whether there might be a greater pressure to produce significant research in dissertations.

The article 'Estimating the reproducibility of psychological science' (Open Science Collaboration, 2015) has already been cited 921 times, which shows that the

subject finds enormous response. However, it is important that the optimisation of research really takes place in practice. For instance, it is still observable that the effect size (Kühberger et al., 2014) and the probability of reproducibility do not get the appropriate attention (Open Science Collaboration, 2015).

The results of the present work suggest that the university as an institution provides a good statistical preparation and culture in the field of economic psychology. The biggest challenge for students seems to be reaching large and representative samples, a problem that manifests in results marked by low statistical power. What this leads to is a necessary consideration of the importance of ensuring that data and statistics are transparent. Because without transparency, no replication and no meaningful interpretation is possible. Indeed, Lakens (2014a) states that ‘Understanding 20% of statistics will improve 80% of your inferences’. The Internet offers useful technological platforms to enable this level of transparency. Moreover, scientists can use Internet platforms to discuss their results, preventing misunderstandings and misinterpretations.

This is not to say that using the Internet to facilitate scientific and statistical transparency is not without risks. The existence of online calculators, which are not scientifically correct, shows that Internet platforms can be incorrect. However, when it comes to ensuring transparency and facilitating rigorous peer review, Internet platforms - like the Open Science Framework, or the signing of ‘A voluntary commitment to research transparency’ (Schönbrodt, 2015) - provide a strong foundation for future research.

9. References

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality, 27*(2), 108-119. doi:10.1002/per.1919

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science. *Circulation research*, *116*(1), 116-126.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to Meta-Analysis, Chapter 7: Converting Among Effect Sizes . Chichester, West Sussex, UK: Wiley.

Bruns, S. B., & Ioannidis, J. P. (2016). P-Curve and p-Hacking in Observational Research. *Plos One, 11*(2). doi:10.1371/journal.pone.0149144

Cohen, B. (2008). *Explaining psychological statistics (3rd ed.)*. New York: John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2. Auflage)*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145-153. doi:10.1037/h0045186

Comte, A. (1830). Cours de philosophie positive. Paris: Bachelier

Elis, P. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge: Cambridge University Press.

Elisabeth, A., Veer, V., & Giner-Sorolla, R. (2016). Journal of Experimental Social Psychology Pre-registration in social psychology – A discussion and suggested template. *Journal of Experimental Social Psychology, 67,* 2-12.

Fanelli, D. (2010). “Positive” Results Increase Down the Hierarchy of the Sciences. *PLoS ONE, 5*(4). doi:10.1371/journal.pone.0010068

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141(1)*, 2.

Fuchs, H. M., Jenny, M., & Fiedler, S. (2012). Psychologists Are Open to Change, yet Wary of Rules. *Perspectives on Psychological Science, 7*(6), 639-642. doi:10.1177/1745691612459521

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology, 13*(3). doi:10.137/journal.pbio.1002106

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science, 23*(5), 524-532. doi:10.1177/0956797611430953

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE, 9*(9). doi:10.1371/journal.pone.0105825

Lakens, D. (2014a). Publication Bias in Psychology: Putting Things in Perspective. Retrieved January 09, 2017, from http://daniellakens.blogs-pot.co.at/2014/09/publication-bias-in-psychology-putting.html

Lakens, D. (2014b). What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *The Quarterly Journal of Experimental Psychology, 68*(4), 829-832. doi:10.1080/17470218.2014.982664

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4:863*. doi:10.3389/fpsyg.2013.00863

Lenhard, W., & Lenhard, A. (2016). *Berechnung von Effektstärken*. verfügbar unter: https://www.psychometrica.de/effektstaerke.html. Bibergau: Psychometrica. DOI: 10.13140/RG.2.1.3478.4245

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology, 65*(11), 2271-2279. doi:10.1080/17470218.2012.711335

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7(1)*, 105-125. doi:10.1037//1082-989X.7.1.105

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:10.1126/science.aac4716

Replicability-Index. (n.d.). Retrieved January 28, 2017, from https://replicationindex.wordpress.com/tag/jpsp/

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews. *Annual Review of Psychology, 52(1)*, 59-82. doi:10.1146/annurev.psych.52.1.59

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (231-244). New York, NY: Sage.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641. doi:10.1037//0033-2909.86.3.638

RStudio, Inc. (2013). Easy web applications in R. Retrieved March 21, 2017, from http://www.rstudio.com/shiny/

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90. doi: 10.1037/a0015108

Schönbrodt, F. (2015). A voluntary commitment to research transparency. Retrieved March 21, 2017, from http://www.nicebread.de/a-voluntary-commitment-to-research-transparency/

Simmons, J. (2013). [6] Samples Can't Be Too Large. Retrieved March 21, 2017, from http://datacolada.org/6

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547. doi:10.1037/a0033242

Stangroom, J. (1999). Quick P-Value Calculators. Retrieved February 28, 2017, from http://www.socscistatistics.com/pvalues/Default.aspx

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician, 49*(1), 108. doi:10.2307/2684823

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance--Or Vice Versa. *Journal of the American Statistical Association, 54*(285), 30. doi:10.2307/2282137

Thinking about evidence and vice versa. (n.d.). Retrieved January 09, 2017, from http://datacolada.org/

Ulrich, R., Erdfelder, E., Deutsch, R., Strauß, B., Brüggemann, A., Hannover, B., . . . Rief, W. (2016). Inflation von falsch-positiven Befunden in der psychologischen Forschung. *Psychologische Rundschau, 67*(3), 163-174. doi:10.1026/0033-3042/a000296

9.1. List of Figures

9.2. List of Tables

10. Appendix

10.1. Abstract in German

Einige Wissenschaftler fanden über die Untersuchung statistischer Kennwerte in psychologischen Studien, über Replikationsversuche und qualitative Erhebungen Hinweise darauf, dass Daten verzerrt dargestellt werden. Basierend darauf war das Ziel dieser Arbeit insbesondere zwei dieser identifizierten Phänomene zu erforschen. Zum einen den bekannten *Publication Bias*, der dafür steht, dass bevorzugt signifikante Ergebnisse berichtet und veröffentlicht werden, bzw. das *File drawer problem*, das beschreibt, dass nicht signifikante Ergebnisse in der Schublade landen, und zum anderen das Phänomen *p-hacking*, welches die Manipulation von ganzen Studien oder einzelnen Analysen bezeichnet, um dadurch unter die berühmte Signifikanzgrenze von .05 zu gelangen.

Die im Vordergrund stehende Erklärung dieser Phänomene und Datenverzerrung geht darauf zurück, dass vorzugsweise interessante und signifikante Ergebnisse veröffentlicht werden und Wissenschaftler aufgrund dessen einem hohen Publikationsdruck ausgesetzt sind.

Die Forschungsfrage möchte daher untersuchen, ob auch Studierenden diese sogenannten „Fragwürdigen Forschungspraktiken“ (*questionable research practices)* anwenden. Dafür wurde in dieser Masterarbeit die statistische Berichterstattung in 254 studentischen Abschlussarbeiten aus dem Bereich der Wirtschaftspsychologie überprüft. Aus einer Abschlussarbeit wurden insgesamt 74 Merkmale mit Hilfe eines Kodierungsschemas erhoben. Dessen Auswertung der statistischen Kennwerte (*p*-Werte, Effektgrößen, Power,…) ergab, dass die Phänomene, wie angenommen aufgrund des fehlenden Publikationsdrucks, hier nicht auftraten. Die Resultate sprechen dafür, dass Studenten auch nicht signifikante Ergebnisse berichten und keine Manipulationen zur Signifikanzerreichung anwenden.

Suchbegriffe: Publication Bias, File drawer problem, *p*-hacking, Signifikanz, statistische Berichterstattung, questionable research practice

10.2. Abstract in English

Scientists give evidence with the help of applying analyses of statistical data in psychological studies, replication experiments and qualitative surveys that psychological data is reported distorted. Based on this, the aim of this study was to explore two of their identified phenomena. On the one hand the well-known *publication bias*, which states that preferably significant results are reported and published, or respectively the *file drawer problem*, which describes the fact that non-significant results remain in the drawer, and on the other hand the phenomenon *p-hacking*, which refers to the manipulation of whole studies or individual analyses in order to get under the well-known significance limit of .05.

The dominant explanation of the phenomena and occurring data distortion is that only interesting and significant results are getting published, and therefore, scientists are exposed to a high publication pressure.

Therefore, the research question would like to examine whether these so-called *questionable research practices* are also applied among students. In this master thesis, the statistical reporting of 254 student theses from the field of economic psychology were examined. With a coding scheme, a total of 74 characteristics were coded out of one thesis. Its evaluation of the statistical characteristic values (*p*-values, effect variables, power ...) showed that the phenomena did not occur, as assumed due to the suggested lack of publication pressure. The results suggest that students do report non-significant results as well as significant results and do not use manipulations to achieve the significance limit.

Keywords: publication bias, file drawer problem, *p*-hacking, significance, statistical reporting, questionable research practice

10.3. 74 statistical characteristics

No.	*Label*	*Explanation*
1	Study code	To anonymize the student thesis
2	Year	In which the thesis was submitted
3	Total page number	To the last page of the literature
4	Study area	Topic of thesis (as a text)
5	Classification	In the decision tree (non-empirical, descriptive…)
6	Hypothesis	Main and salient or first hypothesis of the thesis
7	Hypothesis page number	For orientation
8	Dependent variable	DV of the main hypothesis
9	Independent variable	IV of the main hypothesis
10	Results start	Printed page number, where results start
11	Results end	Printed page number, where results end
12	Sample size	Sample size of tested main hypothesis
13	Sample size page number	For orientation
14	Sample size criterion	Was the sample size justified (power analysis)?
15	Sample size type	Type of sample (psychology students, adults…)
16	Age mean	Overall mean of age of the sample
17	Age standard deviation	Overall standard deviation (SD) of age of the sample
18	Analysis	Which analysis was performed?
19	Analysis page number	For orientation
20	*p*-value type	How was the *p*-value reported (exactly 3 digits…)?
21	*p*-value	*p*-value information from text
22	*p*-value page number	For orientation
23	Effect size hypothesis	Was an effect size reported in the relevant test?
24	Effect size general	Were effect sizes generally reported, when sign.?
25	Effect size type	Which type of effect size was reported?
26	Effect size	Actual effect size
27	Covariates	Were there third variables in the model?
28	Which covariate	Which (as a text)

29	Chi-square-value	Reported empirical chi-square-value
30	Chi-df	Degrees of freedom (df) of chi-square-value
31	t-value	Reported empirical t-value
32	t-df	df of t-value
33	t-type	Was the test independent or dependent?
34	t-correlation	When dependent, was the correlation reported?
35	F-value	Reported empirical F-value
36	F-df reported	Were the df reported along with the F-value?
37	F-df1	df 1
38	F-df2	df 2
39	F-group number	Number of groups that were compared
40	r-value	Reported empirical r-value
41	Regression stand. beta	Reported standardized beta-value
42	Regression unstand. beta	Reported unstandardized beta-value
43	Reg. variance of DV	When unstandardized, variance of DV
44	Reg. variance of IV	When unstandardized, variance of IV
45	z-value	Reported empirical z-value
46	Non-parametric type	Which type of non-parametric analysis was used?
47	Non-parametric value	Reported empirical test value
48	Non-parametric df	df
49	Group 1 size	Sample size of group 1
50	Group 1 mean	Mean value of DV of group 1
51	Group 1 SD	SD of DV of group 1
52	Group 2 size	Sample size of group 2
53	Group 2 mean	Mean value of DV of group 2
54	Group 2 SD	SD of DV of group 2
55	Group 3 size	Sample size of group 3
56	Group 3 mean	Mean value of DV of group 3
57	Group 3 SD	SD of DV of group 3
58	Group 4 size	Sample size of group 4
59	Group 4 mean	Mean value of DV of group 4
60	Group 4 SD	SD of DV of group 4
61	Group 5 size	Sample size of group 5

62	Group 5 mean	Mean value of DV of group 5
63	Group 5 SD	SD of DV of group 5
64	Reporting	What was missing in the reporting?
65	Data used already	Was the data reported in another thesis already?
66	Data same main hypo.	Was the main hypothesis the same?
67	Data which code	Which thesis used the same data (study code)?
68	Comment	General comment (as a text)
69	Intercorrelation	Intercorrelation between IV and DV
70	Inter. page number	For orientation
71	Inter. *p*-value type	*p*-value type of intercorrelation
72	Inter. *p*-value	*p*-value of intercorrelation
73	Exploration	Was there an exploration section?
74	Further testing	Was there a further testing beyond hypo. testing?