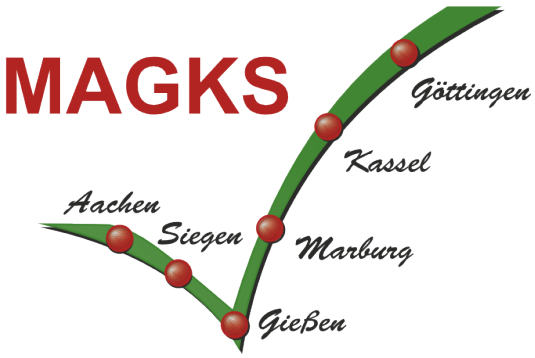

Joint Discussion Paper Series in Economics

by the Universities of
Aachen • Gießen • Göttingen
Kassel • Marburg • Siegen

ISSN 1867-3678

No. 04-2025

Anastasiya-Mariya Asanov, Igor Asanov, Guido Buenstorf, Valon Kadriu, and Pia Schoch

The Origins of Reporting Bias: Selective but Unbiased Reporting by Early-Career Researchers?

This paper can be downloaded from:

https://www.uni-marburg.de/en/fb02/research-groups/economics/macroeconomics/research/magks-joint-discussion-papers-in-economics

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

The Origins of Reporting Bias: Selective but Unbiased Reporting by Early-Career Researchers?

Authors

Anastasiya-Mariya Asanov (ORCID: 0000-0003-3080-4213; noha@incher.uni-kassel.de) [1]

Igor Asanov (ORCID: 0000-0002-8091-4130; igor.asanov@uni-kassel.de) [1,*]

Guido Buenstorf (ORCID: 0000-0002-2957-5532; buenstorf@uni-kassel.de) [1]

Valon Kadriu (ORCID: 0009-0006-1101-5349; kadriu@uni-kassel.de) [1]

Pia Schoch (ORCID: 0009-0006-9471-4590; p.schoch@uni-kassel.de) [1]

[1] University of Kassel, INCHER and Institute of Economics, Kassel, Germany

[*] Corresponding author (e-mail: igor.asanov@uni-kassel.de)

Abstract

Doctoral dissertations provide evidence about research practices in early-stage research. We examine reporting bias by manually collecting over 94,000 test statistics from a random sample of German dissertations and their follow-up papers worldwide. We observe selective reporting, as only a fraction of the tests in the dissertations is reported in follow-up papers. Unexpectedly, we find no increase in reporting bias in follow-up papers compared to dissertations nor, generally, reporting bias in dissertations or papers. Self-selection into higher-impact journals based on statistical significance may reconcile our finding of selective yet "unbiased" reporting with prior evidence suggesting pervasive reporting bias.

Key words: research transparency, reporting bias, higher education, young researchers

JEL codes: A14, A23, C12, I23

Funding acknowledgement: All authors gratefully acknowledge funding by the German Federal Ministry of Education and Research (BMBF) under Grant number 16PH20011 (PI: GB) and via the German Competence Network for Bibliometrics under Grant number 16WIK2101A. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ensuring the integrity and transparency of research is paramount. However, pervasive issues such as publication bias and *p*-hacking present significant challenges to the reliability and validity of scientific literature. Publication bias, characterized by the preferential publication of positive results while neglecting null findings, undermines evidence-based decision-making processes across disciplines (Bruns and Kalthaus 2020). Driven by the academic reward system that prioritizes novelty over rigor, publication bias perpetuates the "file drawer problem", where non-significant findings remain unpublished (Franco, Malhotra, and Simonovits 2014). Furthermore, *p*-hacking exacerbates this bias, as researchers manipulate data and analysis to produce statistically significant results (Brodeur et al. 2023).

Various studies have analysed the prevalence of reporting bias in different fields and settings (Askarov et al. 2023; Bruns et al. 2024; Brodeur, Cook, and Heyes 2020; Brodeur et al. 2023; Brodeur, Cook, and Neisser 2024; Gerber and Malhotra 2008a; 2008b; Vivalt 2019; Ioannidis, Stanley, and Doucouliagos 2017; Franco, Malhotra, and Simonovits 2014; Doucouliagos, Hinz, and Zigova 2022; Brodeur, Kattan, and Musumeci 2024; Chopra et al. 2023). Building on this literature, we trace reporting bias in a large representative sample of PhD dissertations and the (published) papers resulting from these dissertations. This enables us to shed light on the reporting behavior of early-career researchers during their doctoral training and on the pathway of reporting results from the dissertation stage to publication.

By providing new evidence on how reporting bias develops along the publication process, our study contributes to the literature on systematic determinants of selective reporting and publication bias. Closely related, Franco, Malhotra, and Simonovits (2014) find that studies from survey-based experiments in the social sciences with strong results (all or most hypotheses supported by statistical tests) are more likely to get written up and published. In contrast, Brodeur et al. (2023), analyzing submitted and accepted manuscripts from the Journal of Human Resources, show that marginally significant results are more likely to be desk-rejected, and that the review process only has little effect on reporting bias.

Studying differences in the reporting behavior of researchers by academic age, Doucouliagos, Hinz, and Zigova (2022) find that reporting bias in the aid effectiveness literature is more prevalent for non-tenured authors and the degree of reporting bias increases with post-PhD academic age for the non-tenured group. Brodeur, Kattan, and Musumeci (2024) show reporting bias in the job market papers of PhD candidates and that just-significant results

in the job market papers are associated with a higher likelihood of academic placement, especially for a position as assistant professor.

Our study adds to the research on reporting bias at early stages of research careers by studying over 94,000 manually collected test statistics from a representative random sample of empirical dissertations defended in Germany in economics, political science, and sociology. We analyze how reporting bias in these dissertations is associated with institutional quality assurance measures. While other studies in the reporting bias literature have considered the role of institutions (Brodeur, Cook, and Heyes 2020), we go a step deeper and consider institutional factors that might play a role during the early career research phase. By matching dissertations with their follow-up papers, we show how the reporting of empirical results unfolds over the publication process.

Drawing upon established methodologies in the literature, such as *z*-curves for graphical inspection (Askarov et al. 2023; Brodeur et al. 2023), binomial tests (I. Asanov, Bühren, and Zacharodimou 2020; Brodeur, Cook, and Heyes 2020; Vivalt 2019; Gerber and Malhotra 2008a; 2008b), the share of statistically significant results (Blanco-Perez and Brodeur 2020), and caliper regressions (Brodeur et al. 2023; Brodeur, Cook, and Heyes 2020), we systematically study reporting bias. Our approach is novel in that we disentangle possible reporting bias by looking at it from three perspectives: the number of overall tests, the overall share of statistically significant results, and statistical significance inside the caliper around commonly checked threshold levels of significance. We assume that changes in these three outcomes can co-occur, so we isolate them by analyzing them separately in a systematic way.

We establish three results for our sample of dissertations and follow-up papers. First, binomial tests do not indicate statistically significant discontinuities at commonly checked levels of significance in dissertations or papers, suggesting a focus on methodological accuracy over publication pressures among PhD candidates. The absence of reporting bias at the dissertation level might be due to PhD students not knowing yet whether they want to continue in academia, so they might not yet feel the competitive pressure of publishing. The observed lack of significant reporting bias at the paper level may be attributed to minimal revisions between dissertations and published versions of subsequent papers, or to the review process effectively identifying and addressing instances of *p*-hacking (Brodeur et al. 2023). Second, in our regression analyses we see that certain institutional quality assurance measures such as graduate schools and mandatory supervision agendas mostly have a negative relationship with

reporting bias. These results highlight the possible influence of institutional environments on research integrity. Third, we find considerable selective reporting in the number of tests as results move from the dissertation stage to published papers. However, we do not find that the probability of reporting bias increases from dissertation to paper. We run several robustness checks focusing only on the main tests (i.e., excluding intercepts, correlations, obvious controls, robustness checks, and appendices), testing different levels of significance, coming to the same conclusion.

We go on to examine possible mechanisms that can explain our unexpected results of selective yet "unbiased" reporting from the dissertation stage to the published paper. To do so, we first consider the timing of publication of follow-up papers. We observe reporting bias in follow-up papers published after the PhD defense, indicating reporting bias when researchers have left the doctoral training stage and possibly moved to a different institutional context, but not before the defense. However, this does not fully explain the absence of reporting bias in published papers as compared to dissertations. We therefore examine possible associations between reporting bias and the impact factor of the journals where follow-up papers are published. This is feasible because we have a representative sample of dissertations and their follow-up papers, which are published in a diverse set of journals. We find a positive relationship between the journal impact factor indicators and the share of statistically significant results in follow-up papers resulting from dissertations. Observed (self-)selection by impact factor on statistical significance in journals with higher impact factor reconciles our finding of "unbiased" selective reporting of results from a representative sample of dissertations with biased selective reporting observed in competitive journals (which are the common focus of prior studies).

I. Data and Methods

A. *Pre-Analysis Plan*

Considering the importance of pre-analysis plans to remedy reporting bias (Brodeur et al. 2024; Imbens 2021), we wrote a pre-analysis plan before collecting the data and conducting the analysis. In the pre-analysis plan, we rationalized the sample size and sampling strategy, and described data sources and data collection protocols. Moreover, we pre-specified the hypotheses to be tested and respective empirical strategies. We also specified a list of control variables. The pre-analysis plan of this paper is available at the Open Science Framework: osf.io/eyqh2. A replication package will be available upon the acceptance of the manuscript.

B. *Explanatory Variables & Hypotheses*

Our analyses are motivated by the conjecture that the reporting of empirical results by early-career researchers is conditioned by quality assurance measures adopted by the university where they do their dissertation research, as well as the process in which dissertation results are turned into published journal papers. In the pre-analysis plan, we formulated the following three hypotheses:

Hypothesis 1 Graduate schools are associated with the reporting of statistical tests.

Graduate schools are one of numerous institutional measures to ensure quality. As mentioned in Section 1, quality assurance measures like graduate schools should ingrain high-quality research practices to ensure the knowledge created is reliable (Hüther and Krücken 2018). Specifically, this can take the form of workshops on methodology or other aspects and presentations of research projects that are discussed with other PhD students and senior researchers of the department.

Hypothesis 2 Mandatory supervision agendas are associated with the reporting of statistical tests.

Mandatory supervision agendas are a relatively new form of quality assurance. The goal is to create a regular exchange of information about the dissertation progress between the PhD student and their supervisor and resolve possible conflicts. These mandatory supervision agendas are primarily structured with pre-specified talking points. Roebken (2007) states that structured formats can be beneficial but unfavorable, depending on the individual project and person.

Hypothesis 3 The publication process is associated with the reporting of statistical tests of dissertations and follow-up articles. [1]

While many studies exist on the reporting of statistical tests, i.e., reporting bias, to our knowledge, there are no studies yet that illuminate reporting bias from the dissertation stage to the paper stage. The closest studies to what we are analyzing are Franco, Malhotra, and Simonovits (2014) and Brodeur et al. (2023). The first study uses 221 published and unpublished studies that were all part of the Time-sharing Experiments in the Social Sciences (TESS). Here, researchers proposed survey-based experiments. These proposals underwent a

[1] In the PAP, Hypothesis 3 dealt with the Handelsblatt Ranking, but we decided to move that to the Online Appendix because we found a very low variability between fields that we could not anticipate in advance, meaning that the Hypothesis 4 from the PAP is now Hypothesis 3.

peer-review process, based on which the researchers could receive a grant to conduct the experiment. However, even though this peer-review process only allowed high-quality experiments to be accepted, the studies with statistically significant results were still more likely to be published and written up (Franco, Malhotra, and Simonovits 2014). Brodeur et al. (2023) use test statistics from manuscripts during the Journal of Human Resources peer-review process to analyze reporting bias. They find considerable discontinuities around the conventional significance thresholds in the initial submissions with fewer discontinuities for manuscripts submitted for peer review compared to the desk-rejected manuscripts, indicating that reporting bias might not be corroborated by the peer-review process but stem from the author side. Lastly, since we are considering dissertations and follow-up papers, our paper also relates to literature that evaluates the association between career stages and reporting bias. H. Doucouliagos, Hinz, and Zigova (2022) use data from studies in the aid effectiveness literature, and they collect data on the authors' career stages, i.e. tenure and age. The authors indicate that the tendency for reporting bias might increase with age, especially for non-tenured researchers (Doucouliagos, Hinz, and Zigova 2022).

C. *Sampling Strategy*

We drew a random sample of 3,000 dissertations from the German National Library's database of dissertations defended at German universities in 2004-2006 and 2012-2014 from dissertations classified in the database under the fields of economy, sociology, and political science.[2] As the economy field in the database comprises dissertations in economics and management, we manually differentiate between them. This resulted in 1,840 dissertations from 73 German universities in economics, sociology, and political sciences, among which we only consider empirical dissertations. Before collecting and analyzing the data and anticipating that only part of the dissertations would be empirical, we ran conservative power calculations for each of our Hypotheses to rationalize our sample size, which we documented in the pre-analysis plan. Below, we provide estimates of the minimum detectable effect size given observed distributional characteristics.

In Hypotheses 1 and 2, we analyze the effect of quality assurance measures of universities on reporting bias in empirical dissertations, i.e., graduate schools and mandatory

[2] Specifically, we drew 1500 doctoral dissertations from 2004 to 2006 and 1500 doctoral dissertations from 2012 to 2014 classified under following three fields in database: "Wirtschaft", "Sozialwissenschaften, Soziologie, Anthropologie", "Politik". We refer to field "Sozialwissenschaften, Soziologie, Anthropologie" as sociology in text.

supervision agendas, so we expect an effect at the university level with a corresponding adjustment of clustered standard errors at the university level (see Abadie et al. (2023). Given 75 tests on average per cluster inside the 0.150 caliper, with 58 observed universities (clusters) and observed intra-university (-cluster) correlation (ICC) of 0.003 for the main outcomes, assuming a conventional significance level of 5 percent and 80 percent of statistical power, we have a minimum detectable effect (MDE) size of 0.1 of a standard deviation. Correspondingly, we are able to detect a 5 percentage points difference (Cohen's H=0.1) from a baseline of 50 percent with an observed ICC of 0.003, 75 tests on average per cluster from 58 universities at a significance level of 5 percent and 80 percent of statistical power.

In Hypothesis 3, we analyze the difference in reporting bias between empirical dissertations and their follow-up papers, so we expect an effect at the dissertation-paper pair level, i.e., at the author level - sampling and treatment level (see Abadie et al. (2023)). Given 331 dissertation-paper pairs (clusters) with 18 tests on average per cluster inside the 0.150 caliper with observed intra-dissertation (-cluster) correlation (ICC) of 0.024 for the main outcomes, assuming a conventional significance level of 5 percent and 80 percent of statistical power, we have the minimum detectable effect (MDE) size 0.09 of a standard deviation. Correspondingly, we are able to detect a 4.4 percentage point difference (Cohen's H=0.9) from a baseline of 50 percent with an observed ICC of 0.024, 18 tests on average per cluster from 331 dissertation-paper (clusters) pairs at a significance level of 5 percent and 80 percent of statistical power.

This is a conservative estimate as the actual number of tests is larger when we test the whole distribution of tests, when we use all tests collected, or when, in addition, we account for the variance explained by covariates. That is, we have sufficient power to detect a correlation of even 0.1 (r), which can be considered small given observed effect sizes in economic research (Ioannidis et al., 2017).

D. *Outcome variables*

The three outcome variables in our paper are (1) the number of test statistics, (2) the share of statistically significant test statistics, represented by an indicator variable for statistical significance at the 5 percent significance level, and (3) the statistical significance at the 5 percent significance level inside a narrow caliper represented by an indicator variable taking the value one if the test statistic is significant at the 5 percent level and zero if not[3]. We focus

[3] In the PAP, we intended to test "difference of the observed *z*-value distribution to the *z*-value distribution under the assumption of no p-hacking" following the approach of a working paper version of Bruns et al. (2024).

on the 5 percent significance level, where one might assume a higher probability of reporting bias. Yet, we additionally run robustness checks for other conventional levels of significance - 1 percent and 10 percent significance.

We calculate the number of test statistics per dissertation and per follow-up paper. The variables for statistical significance are calculated based on the test statistics reported in the dissertations and the papers. We convert reported test statistics into *z*-values using the following hierarchy: If *p*-values are reported, we convert them into z-values; if *t*-values are reported, we treat them as *z*-values; and if *z*-values are reported, we use them as they are. However, if a coefficient and a standard error are reported and there is no *p*-value, *t*-value, or *z*-value, we calculate the *z*-value based on them by dividing the coefficient by the standard error. Also, we extract test statistics precisely as reported in the manuscripts. For cases where only the coefficient and standard errors are reported, we follow the approach of Kranz and Pütz (2022) to remove imprecisely reported test statistics. In the following paragraphs, we explain in more detail how we collected the relevant data.

Dissertation Data

Before collecting the test statistics of each dissertation, we had to obtain the dissertations and their contents first. We have, in total, 1,840 dissertations in the fields of economics, sociology, and political science. The German National Library (DNB) database was used to collect dissertation- and author-level data, mainly to construct control variables. A complete list can be found in the **Online Appendix Table A**. In our random sample, we consider dissertations in any format, meaning that due to most dissertations not being available digitally (around 1,500), we coded them manually to prevent selection bias.[4] This laborious process was done by four research assistants who strictly followed a protocol created by two senior researchers. The manual coding took around 30 minutes per dissertation. After coding the dissertations, we constructed a manual algorithm to extract all regression tables from them.[5] Because our sample includes dissertations in German and English, we used keywords in both languages indicating regression tables, i.e., “regression,” “OLS,” “logit,” “significance,” “t-value,” “t-Wert,” “standard error,” “Standardfehler,” “coefficient,” and “Koeffizient.”

However, it turned out that this method was not applicable to our data as it is meant for meta-analyses. As a substitute method, (a) we apply the standard Kolmogorov-Smirnov test (KS test) to measure the difference between the distributions in question and (b) in regressions, assess the difference in the distribution of statistical significance inside a caliper as mentioned in the text.

[5] While laborious, manual data collection allowed us to ensure data quality. In addition, see benefits of human teams in error detection for assessing research reproducibility, compared to innovative AI-based approaches (Brodeur et al. 2025)

This algorithm allowed us to identify empirical dissertations and extract the relevant pages, including test statistics. 327 dissertations are empirical, reporting coefficients, standard errors, *t*-values, *z*-values, or *p*-values. These 327 dissertations were randomly assigned to two of the authors of this paper. When extracting all test statistics from the dissertations, the coders strictly followed the data collection protocol pre-specified by the team. This includes the test statistics mentioned above and the reported significance by means of eye-catchers, the number of observations of each model, an indication if the test statistic was used in a two-sided test or not, and if the test statistic originates from the main analysis or a robustness check. We define robustness checks as analyses in the main part of the dissertation that explicitly mention the words "robustness check" or "sensitivity analysis" in the table header or any analyses located in the Appendix. We also extracted information on the data source (own data, external data), data type (cross-section, panel, time series), general research design (lab experiment, field experiment, quasi-experiment, observational), and research design sub-categories (randomized controlled trial, instrumental variables, regression discontinuity, difference-in-difference,).

We extracted regression coefficients, standard errors, *t*-values, *z*-values, *p*-values, and correlation coefficients from every table and every model because we assume that all of these can be used, and are being used, in the research community to make claims about research findings. To ensure high quality of the data, two other researchers (a senior researcher and a research assistant) drew a 5 percent random sample of the test statistics and checked for correctness. Systematic errors or uncertainties were documented and sent to the initial coders for checking. Based on these comments, the initial coders went through the whole dataset again, cleaning up systematic errors raised by the third senior researcher and the research assistant.

Paper Data

We applied the exact same approach for the follow-up papers. To identify follow-up papers of dissertations, we compared the titles, abstracts, and, if needed, the introductions of each dissertation with each of the authors' articles. The quality of this approach was ensured by two methods. Firstly, we encouraged coders to leave comments in case of uncertainties, which were resolved in deliberation with the senior researchers. Secondly, we drew a 30 percent random sample and requested two different team members to apply the same approach as the initial coders independently and without knowing their results. This approach resulted in a 94 percent agreement between the initial coders and the validators. For details regarding the matching of dissertations and their follow-up papers, see A.-M. Asanov et al. (2024).

A total of 301 empirical articles were identified as follow-up papers of the dissertations. These were randomly assigned to six research assistants, who were given identical instructions to collect test statistics as described in the previous subsection to ensure the uniformity in data collection between the dissertation and article statistics. We intentionally assigned the paper data to be collected by different employees to prevent possible bias from knowing dissertations' distribution of test statistics. Research assistants then independently cross-checked each other's data for systematic errors in the same manner as was done for the dissertations (a 5 percent random sample was created for each batch of coded papers for cross-checking), which the initial coders then considered in the first cleaning step. To further ensure high quality of the data, one of the senior researchers checked the collected data on the follow-up papers for possible errors, which were then considered by the research assistants in a final data cleaning step.

Moreover, to be close to the literature that considers solely "hypothesis-testing" test statistics, we identified obvious control variables in the dissertation and paper data. To do so, three authors of this paper and a research assistant were randomly assigned dissertations and follow-up studies to identify obvious controls. They identified coefficients explicitly labeled either in the regression table or in the text as control variables. Using this approach, we aimed to ensure that identifying control variables is as objective as possible. Lastly, the identification of chapters in cumulative (i.e., paper-based rather than monographic) dissertations that later got published was carried out by two research assistants who both received the same dataset. To ensure that the chapters were categorized correctly, one of the authors examined their results for possible discrepancies and resolved them. See Figure 1 for a detailed diagram of the data collection process.

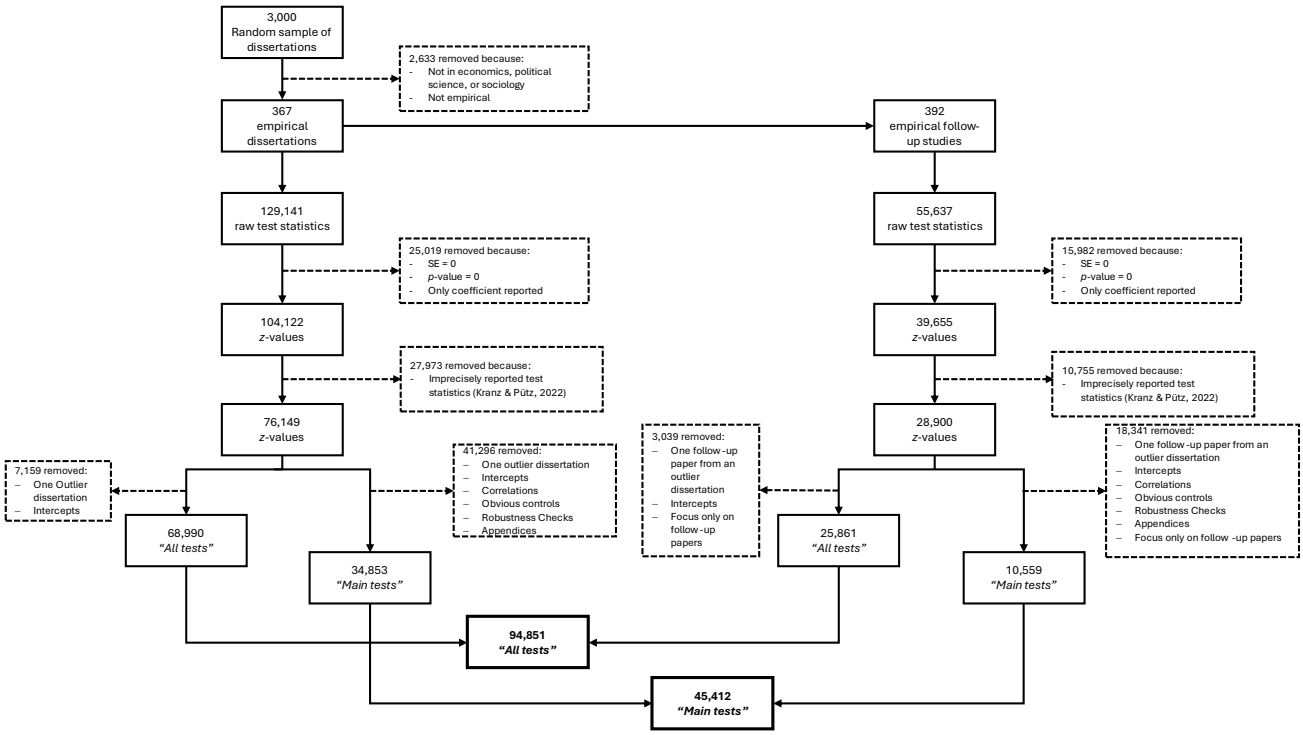

Figure 1. CONSORT Diagram showing the data collection process for the dissertations on the left side and for their follow-up publications on the right side. The two boxes at the end show the overall number of test statistics when combining dissertations and follow-up papers.

E. *Empirical Strategy*

To test the hypotheses, we apply various methods that we explain in the following paragraphs. In the first step, we inspect our data by visualizing the distribution of *z*-values with density curves (Brodeur, Cook, and Heyes 2020; Brodeur et al. 2023). This allows us to find clues for reporting bias by inspecting possible discontinuities around the common thresholds for statistical significance, as well as general changes in z-value distributions between dissertations and follow-up papers.

In the second step, we apply binomial tests to analyze the *z*-value distribution analytically in small windows around the critical *z*-values (I. Asanov, Bühren, and Zacharodimou 2020; Brodeur, Cook, and Heyes 2020; Gerber and Malhotra 2008a; 2008b; Vivalt 2019). W apply three different calipers to our data: 0.150, limiting our data to $z \in [1.81, 2.11]$; 0.050, limiting our data to $z \in [1.91, 2.01]$; and 0.010, limiting our data to $z \in [1.95, 1.97]$. A success is defined as the test statistic being statistically significant at the 5 percent level, and the number of trials is the number of test statistics inside the caliper. We assume a

binomial probability of 0.5 because, with no reporting bias, one would expect as many test statistics just under the critical threshold as just above the critical threshold. Because our data resembles a panel structure, we account for the non-independence of test statistics within dissertations/papers by applying bootstrapping on the author level following the approach of I. Asanov, Bühren, and Zacharodimou (2020). We report 95 percent confidence intervals.

In the final step, we run different regressions that take the form of the following pre-specified functional form:

$$Y_{ijk} = f(\beta_0 + \beta_{GS} GS_{ijk} + \beta_{SA} SA_{ijk} + \beta_{Paper} Paper_{ijk} + X_{ijk}) \quad (1),$$

where *i* is dissertation, *j* is university, and the index *k* stands for the test statistic. Y_{ijk} is one of the following outcome variables: the number of test statistics, an indicator variable taking the value one if the test statistic is statistically significant at least at the 5 percent level and zero if not, or an indicator variable taking the value one if the test statistic is statistically significant at least at the 5 percent level inside the caliper and zero if not.

The variable GS_{ijk} stands for graduate school and takes the value one if a graduate school in the field of the dissertation was active at the university before completion of the dissertation and zero if not. To ensure comparability, we only consider graduate schools that received German Research Foundation (DFG) funding either directly (Graduiertenkolleg) or through the excellence initiative. The variable SA_{ijk} stands for mandatory supervision agenda and takes the value one if signing a supervision agenda was required at the university before the completion of the dissertation and zero if not. The variable $Paper_{ijk}$ takes the value one if the test statistic is from a follow-up paper and 0 if it is from a dissertation. X_{ijk} is a vector of control variables on the author, dissertation, and university levels. Controls are selected through the post-double-selection Lasso procedure (Belloni, Chernozhukov, and Hansen 2014) from the pre-specified list of controls (see **Online Appendix Table A**). We consider control variables with less than 20 percent missing values (Cilliers, Elashmawy, and McKenzie 2024). For Hypotheses 1 and 2, where we test university characteristics, we cluster standard errors by universities because we hypothesize an effect on the university level (Abadie et al. 2023). For Hypothesis 3 on publishing process, we cluster standard errors by authors since we hypothesize an effect on the author level and in line with the sampling design (Abadie et al. 2023).

To our knowledge, our estimation strategy is innovative in detecting reporting bias as it studies distributional characteristics over a broad spectrum of outcomes that can change simultaneously. It consists of three different regressions, which we apply to each of the

hypotheses mentioned in Section IB. Because we assume that a change in the mentioned outcome variables (number of test statistics, share of statistical significance, statistical significance inside the caliper) can occur simultaneously, we disentangle them by analyzing them separately in a systematic way.

We first run an OLS regression with the number of tests as the outcome variable and each variable of interest corresponding to the hypotheses. Then, we run a logit regression[6] with an indicator as the outcome variable, taking the value one if the test statistic is statistically significant at least at the 5 percent level and zero if not. Another logit regression is run with the outcome variable being an indicator variable, taking the value one if the test statistic is statistically significant, at least at the 5 percent level inside a narrow caliper of 0.150. We follow the approach of (Kranz and Pütz 2022) and remove imprecisely reported test statistics before performing regression analysis.

II. Results

A. *Descriptive Analysis*

Our manually collected dissertation-based dataset set consists of 69,990 test statistics from 327 empirical dissertations after a conservative data cleaning procedure, e.g. deleting imprecisely reported test statistics (Kranz and Pütz 2022). **Table 1** reports descriptive statistics of our data. Of the dissertations in the sample, 86 percent are in economics, 11 percent are in sociology, and 3 percent in political science. Fifty-seven percent of the dissertations in our sample are cumulative dissertations, with 68 percent of test statistics originating from these. Dissertations in English comprise 72 percent of our sample (79 percent of test statistics), and 67 percent of the dissertations in our sample were defended after 2007 (74 percent of test statistics). 23 percent of test statistics originate from analyses that used their own data. Regarding the data type, 39 percent of test statistics originate from cross-section analyses, 20 percent from time-series analyses, and 51 percent from panel data analyses. Regarding research design, 4 percent of test statistics originate from analyses using lab experiments, 4 percent from analyses using field experiments, 4 percent from analyses using quasi-experiments, 86 percent from observational analyses, 7.5 percent from analyses using instrumental variables approaches, 0.1 percent from analyses using regression discontinuity design approaches, 4.6

[6] In the PAP, we mentioned that we will run a beta regression. We still report the beta regression results in the **Online Appendix Table D2**, but will stick to the logit regression on test-level in the main analysis due to this approach having more power.

percent from analyses using difference-in-difference approaches, and 0.1 percent from analyses using randomized controlled trials.

Of our 69,990 test statistics, 3 percent directly report the *z*-statistics, 10 percent report p-values, 29 percent report *t*-values, 58 percent report coefficients and standard errors. Robustness checks and appendices account for 29 percent of the test statistics. Regarding the dissertation authors, 31 percent are female, 39 percent have a spouse, and 35 percent have an international education.[7] Regarding the quality assurance measures, 31 percent were written at universities with a graduate school in place, and 14 percent were written at a university with a mandatory supervision agenda.[8]

The paper-based dataset consists of 25,861 test statistics from 301 empirical follow-up papers after cleaning the dataset. **Table 1** reports descriptive statistics. Eighty-eight percent of the follow-up papers originate from economics dissertations, 9 percent from sociology dissertations, and 2 percent from political science dissertations. Regarding the type of dissertation, 82 percent of follow-up papers originate from cumulative dissertations. 86 percent of follow-up papers originate from dissertations written in English, and 87 percent of follow-up papers were published after 2007. Regarding the data sources and type, 15 percent of test statistics originate from analyses using their own data. 30 percent originate from cross-section analyses, 10 percent from time-series analyses, and 64 percent from panel data analyses. Regarding research design, 3 percent of test statistics originate from analyses using lab experiments, 3 percent from analyses using field experiments, 18 percent from analyses using quasi-experiments, 89 percent from observational analyses, 11 percent from instrumental variables approaches, 0.6 percent from regression discontinuity design approaches, 10 percent from difference-in-difference approaches, and 0.08 percent from randomized controlled trials.

[7] The information about spouse based on manually classified acknowledgment section.

[8] The number of dissertations in the subgroups for the type of reporting does not add up to the total number of dissertations because one dissertation might use different types of reporting.

Table 1—Summary Statistics for dissertations and follow-up papers[9]

	Dissertations			Follow-Up Papers		
	%	No. of dissertations	No.of Teststats	%	No. of papers	No.of Teststats
Total	**100%**	**327**	**68,990**	**100%**	**301**	**25,861**
Economics	86%	281	59,955	88%	266	24,479
Sociology	11%	36	7,532	9%	28	1,180
Political Science	3%	10	1,503	2%	7	202
Cumulative Dissertation	57%	185	47,145	82%	246	22,849
Dissertation in English	72%	236	54,305	86%	259	23,828
Post 2007	67%	219	50,806	87%	262	23,572
Data Sources						
Own Data	36%	119	15,708	29%	86	3,818
External Data	71%	233	54,391	74%	222	22,135
Data Type						
Cross-Section	49%	161	26,873	42%	125	7,652
Time Series	21%	69	14,085	13%	40	2,475
Panel	42%	137	34,906	50%	149	16,603
Research Design						
Lab Experiment	8%	25	3,001	7%	20	704
Field Experiment	4%	12	2,639	3%	9	752
Quasi Experiment	3%	11	2,958	13%	38	4,616
Observational	88%	289	59,517	85%	257	22,990
IV	6%	19	5,188	7%	20	2,719
RDD	0.3%	1	78	1%	3	169
DID	4%	14	3,204	7%	22	2,505
RCT	0.3%	1	47	0.3%	1	20
Test Statistics						
z-value	11%	35	1,892	8%	23	1,057
p-value	28%	93	6,567	31%	92	3,626
t-value	33%	109	20,216	24%	72	6,183
Coefficient/Standard Error	58%	191	40,315	60%	180	14,995
Robustness Extension	**39%**	**127**	**19,704**	**35%**	**106**	**8,986**
Author Data						
Female	31%	101	17,250	31%	93	7,215
Spouse	39%	127	22,811	37%	111	7,902
International Education	35%	116	26,558	35%	104	7,671
Quality Assurance Measure Data						
Graduate School	31%	100	21,544	32%	96	7,717
Mandatory Supervision Agenda	14%	45	12,294	20%	60	6,238

Of the 25,861 test statistics in the paper-based dataset, 4 percent directly report the *z*-statistics, 14 percent report *p*-values, 24 percent report *t*-values, 58 percent report coefficients and standard errors. Thirty-five percent of the test statistics originate from robustness checks or appendices. When looking at the dissertation authors, 31 percent are female, 37 percent have

[9] We report *z*-values after the approach of Kranz and Pütz (2022) was applied to remove imprecisely reported test statistics. We only exclude intercepts and one outlier dissertation, which we also exclude throughout the further analyses. Summary statistics for the journal locations can be found in the **Online Appendix Table B**.

a spouse, and 35 percent have an international education. Regarding the quality assurance measures, 32 percent were written at universities with a graduate school in place, and 20 percent were written at a university with a mandatory supervision agenda. The distribution of the journal publishing location of the follow-up papers looks as follows: 75 percent of papers are published in European journals, 21 percent are published in North American journals, 2 percent in South American, 1 percent in Asian journals, and 1 percent published in African journals (see **Online Appendix Table B**).

In the next step, we plot *z*-curves to inspect the distribution of the test statistics graphically. Under no reporting bias, one would expect a *z*-curve without discontinuities around the conventional thresholds for statistical significance. **Figure 2** shows density curves for *z*-values $\in [0, 10]$ with the Gaussian density on the ordinate. The vertical black lines depict the *z*-values corresponding to the conventional thresholds for 10 percent, 5 percent, and 1 percent statistical significance. In Panel A, we consider all tests (i.e., excluding only intercepts), and the distribution of 63,477 tests does not show any visible discontinuity around any of the common significance thresholds, showing no indication of reporting bias. In Panel C, we plot the distribution of 63,477 *z*-values for dissertations (red solid line) and 23,705 *z*-values for the follow-up papers (blue solid line). The distributions are almost identical around the significance thresholds, indicating 10 percent, 5 percent, and 1 percent significance. Considering the distributions to the right side of the 1 percent significance threshold ("larger" *z*-values), it seems that the proportion of these is higher for the papers compared to the dissertations, albeit to a small extent. We see a very similar story when visualizing the same Gaussian density curves for the main tests (i.e., excluding intercepts, correlations, obvious controls, robustness checks, and appendices) of dissertations (32,223 *z*-values) and follow-up papers (9,729 *z*-values). In Online Appendix Figure A1 -A3, we report the Gaussian density curves for *z*-values $\in [0, 5]$ and for the presence/absence of graduate schools and mandatory supervision agendas, respectively.

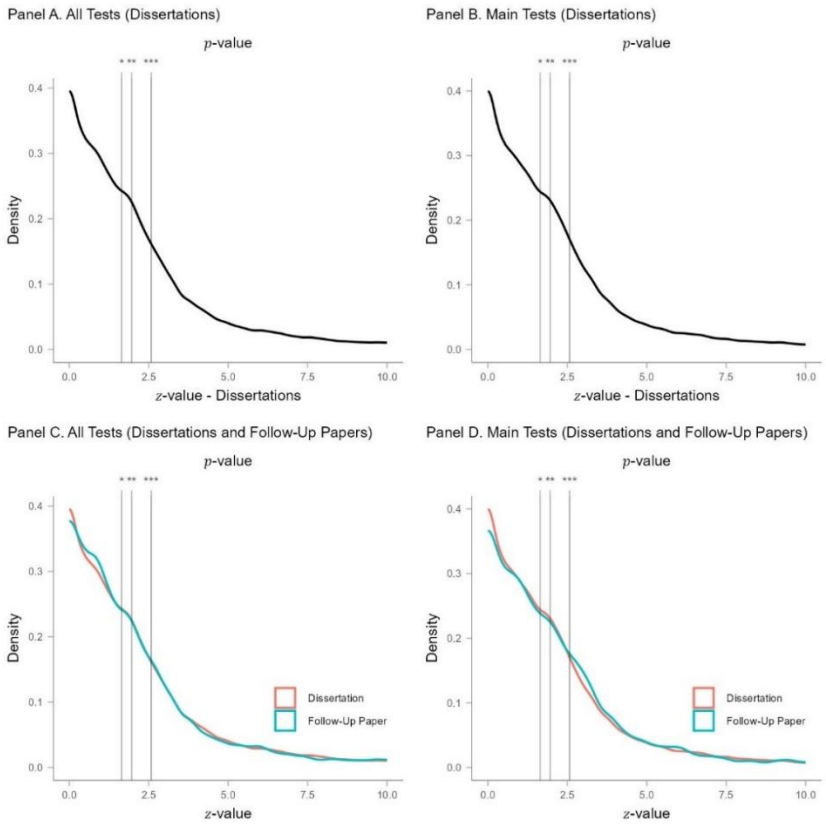

Figure 2. Distribution of *z*-values depicted by Gaussian density curves. In all panels, we consider only z ∈ [0, 10]. Black vertical lines depict *z*-values corresponding to the conventional significance thresholds for 10 percent, 5 percent, and 1 percent. Imprecisely reported z-values removed following (Kranz and Pütz 2022).

We run the Kolmogorov-Smirnov test to check for statistically significant differences between the distributions as pre-specified in the pre-analysis plan. We follow the approach of Brodeur, Cook, and Heyes (2020) and apply the test to the whole distribution and an interval around the critical thresholds. When applying the KS test to the whole data excluding only intercepts, it suggests a difference for the entire distribution (p = 0.0564), but it does not indicate a difference for the intervals around the critical thresholds, i.e., $z \in [1.65, 2.58]$ (p = 0.6277), and $z \in [1.65, 1.96]$ (p = 0.4905). When applying the KS test to the data containing only the main tests (i.e., excluding intercepts, correlations, obvious controls, and tests from robustness checks and appendices), it again suggests a difference for the entire distribution (p = 0.0025), but does not indicate a difference for the intervals around the critical thresholds, i.e., $z \in [1.65, 2.58]$ (p = 0.4953), and $z \in [1.65, 1.96]$ (p = 0.9157). In other words, while we observe hypothesized changes in the distribution, they do not seem to be related to the selective reporting typically found in the literature.

B. *Binomial Tests*

We proceed to a more detailed analysis of discontinuities with the help of a binomial test. In the first two columns, we run the binomial tests on the whole dataset; in columns 3-4, we consider only test statistics from dissertations; in columns 5-6, we consider only test statistics from follow-up papers. Moreover, we apply the binomial tests for the main tests only (i.e., excluding intercepts, correlations, obvious controls, robustness checks, and appendices) and for all tests (i.e., excluding only intercepts) for each caliper and each subset.

In the first panel of **Table 2**, we analyze the subset using a 0.150 caliper. For the overall data in this subset focusing only on main tests, we have 2,952 test statistics, of which 1,453 are statistically significant, and 1,499 are not. The binomial proportion is 0.492, and the lower bound of the 95 percent confidence interval is below 0.5. We find similar results when considering all tests. We still find no signs of reporting bias when focusing on the main tests or when considering all tests when doing the same exercise for the dissertation and follow-up paper subsets separately for the 0.150 caliper.

For the 0.050 caliper, the binomial proportion for the overall data focusing only on main tests is 0.527, and the lower bound of the 95 percent confidence interval is 0.501. We find a similar binomial proportion when considering only main test statistics from dissertations (0.517). In this case, the lower bound of the 95 percent confidence interval is slightly below 0.500. When considering only the follow-up papers, the binomial proportion is 0.561, and the lower bound of the 95 percent confidence interval is exactly 0.500. Considering all tests, we find binomial proportions close to 0.500 for every subset in the 0.050 calipers, but every lower bound of the 95 percent confidence interval is lower than 0.500. These results may suggest a minor extent of reporting bias inside the 0.050 calipers driven by the follow-up papers when considering the main tests only.

When considering only *z*-values inside the smallest caliper, 0.010, the binomial proportion is lower than 0.500 for the overall data with the main tests only and all tests. Looking at the dissertations and follow-up papers separately, we find binomial proportions lower than 0.500, and the lower bounds of the 95 percent confidence intervals are also lower than 0.500, suggesting, surprisingly, no presence of reporting bias both when focusing on the main tests and when considering all tests.

Lastly, we run the same analysis for the 10 percent and 1 percent significance levels, applying the same calipers and excluding only the intercepts (see **Online Appendix Tables C1 and C2**, respectively), and, surprisingly, find no indications of reporting bias for the 10 percent

level, even though normally in the literature, the reporting bias is found at the 10 percent or 5 percent levels. Only at the rarely considered conservative 1 percent level, we detect possible reporting bias for the dissertations and reporting bias for follow-up papers.

Table 2—Binomial Tests for the 5 percent significance level

	All		Dissertation		Follow-Up Paper	
	Main Tests	**All Tests**	**Main Tests**	**All Tests**	**Main Tests**	**All Tests**
Caliper Size			0.150			
No. of Tests in Caliper	2,952	6,015	2,275	4,373	677	1,642
Under Caliper	1,499	3,084	1,168	2,249	331	835
Over Caliper	1,453	2,931	1,107	2,124	346	807
Binomial Probability	0.492	0.487	0.487	0.486	0.511	0.491
95% Confidence Interval	[0.478, 0.507]	[0.476, 0.500]	[0.471, 0.501]	[0.473, 0.499]	[0.479, 0.551]	[0.471, 0.512]
Caliper Size			0.050			
No. of Tests in Caliper	991	2,050	777	1,496	214	554
Under Caliper	469	993	375	734	94	259
Over Caliper	522	1,057	402	762	120	295
Binomial Probability	0.527	0.516	0.517	0.509	0.561	0.532
95% Confidence Interval	[0.501, 0.552]	[0.495, 0.534]	[0.491, 0.544]	[0.490, 0.527]	[0.500, 0.622]	[0.496, 0.574]
Caliper Size			0.010			
No. of Tests in Caliper	211	450	166	321	55	129
Under Caliper	116	245	93	172	23	73
Over Caliper	95	205	73	149	22	56
Binomial Probability	0.450	0.456	0.440	0.464	0.489	0.434
95% Confidence Interval	[0.390, 0.517]	[0.413, 0.502]	[0.379, 0.505]	[0.421, 0.508]	[0.370, 0.630]	[0.344, 0.554]

Notes: In this table, we present the results of binomial proportion tests for test statistics for our overall dataset and for each test statistic from dissertations and their follow-up papers. We also differentiate between "main tests" (i.e. excluding intercepts, correlations, obvious controls, robustness checks, and appendices) and "all tests" (i.e. excluding only intercepts). A success is defined as a statistically significant observation at the 5 percent significance level. In the first panel, we use observations where ($1.81 < z < 2.11$); in the second panel, we use observations where ($1.91 < z < 2.01$); in the last panel we use observations where ($1.95 < z < 1.97$). We then test if this proportion is statistically greater than 0.5. We apply bootstrapping to consider non-independence between observations and report the 95 percent confidence intervals.

C. *Regressions*

In **Table 3.1,** we first report the main regression results on all tests (only excluding the intercepts).[10] Coefficients in Models 1 and 4 are calculated with OLS regression using the aggregated data. In contrast, those in Models 2-3 and 5-6 were calculated using logit regression using the long data and average marginal effects are reported.

First, we look at institutional factors (Models 1-3). For the dissertations, we do not find any statistically significant association between the presence of a DFG-funded graduate school program and selective reporting bias. We also do not see selective reporting in the number of tests or an increase in the share of statistically significant tests in the presence of a mandatory

[10] See **Online Appendix Table D1** for Table 3.1 with control variables displayed.

supervision agenda. The presence of a mandatory supervision agenda, however, is negatively associated with reporting statistically significant results at 5 percent inside the 0.150 caliper (-8.1 percentage points difference).

Next, we assess the evolution of selective reporting and reporting bias from the dissertation to the paper stage (Models 4-6). Considering all dissertations and all follow-up papers, we find that the number of test statistics is reduced from the dissertation to the paper stage. This can be explained by the fact that not all chapters from the dissertation result in a published paper. We, however, do not find an emergence of reporting bias from the dissertation- to the paper stage, despite a change in the number of tests reported. Specifically, we do not see any increase in reporting bias when considering the share of statistically significant tests as the outcome (-1 percentage points difference) nor when considering the indicator variable for statistical significance at 5 percent inside a 0.150 caliper (-3.4 percentage points difference). Both coefficients are negative and not statistically significant.

Table 3.1—Main Regression

	Dissertations			**Dissertations/Papers**		
	Number of Tests (1)	**Share Stat. Sig. 5% (2)**	**Significant at 5% (3)**	**Number of Tests (4)**	**Share Stat. Sig. 5% (5)**	**Significant at 5% (6)**
Graduate School	-26.707	-0.001	0.007			
	(39.225)	(0.009)	(0.028)			
Mandatory Supervision Agenda	104.511	-0.003	-0.081			
	(71.118)	(0.013)	(0.032)			
Paper				-150.037	-0.010	-0.034
				(19.829)	(0.012)	(0.029)
Observations	327	54565	3433	626	72390	5263
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
PDL Controls	Yes	Yes	Yes	Yes	Yes	Yes
Caliper	-	-	[1.81, 2.11]	-	-	[1.81, 2.11]
R^2	0.129	-	-	0.140	-	-
Adj. R^2	0.050	-	-	0.098	-	-
R^2 tjur	-	0.817	0.011	-	0.803	0.020
RMSE	-	0.212	0.497	-	0.221	0.495
F	-	149.65	1.079	-	232.88	0.997

Note: Models 1-3 report regression results from dissertations only. Models 4-6 report regression results from dissertations and follow-up studies. In Models 1 and 4, we apply OLS regression, with the outcome variable being the count of test statistics per dissertation. Models 2-3 and 5-6 report average marginal effects from logit regressions, with the outcome variable in Models 2 and 5 being an indicator variable for at least 5 percent significance and an indicator variable for at least 5 percent significance in Models 3 and 6. Models 3 and 6 consider only test statistics inside a 0.150 caliper around the 1.96 z-value, i.e. absolute z-values between 1.81 and 2.11. Imprecise z-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with Post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Year FE and Region FE were kept fixed. Standard errors are clustered at the university level in Models 1-3 and at the author level in Models 4-6.

Regarding the control variables selected with the help of a machine learning algorithm, we find commonly suspected variables associated with selective reporting bias (see **Online Appendix Table D1**). We find that using *eye-catchers* at different levels of statistical significance is statistically significantly associated with the number and share of statistically significant tests. Another set of controls selected by the machine learning algorithm indicates that *methods matter* in selective reporting bias in line with findings by Brodeur, Cook, and Heyes (2020). Namely, post-double lasso selects variables indicating if the test statistic originates from a difference-in-differences, instrumental variable, regression discontinuity design, or randomized controlled trial analysis, as well as other methodological characteristics. This highlights heterogeneity by method in selective reporting and reporting bias.

Author-level variables are also selected as controls in regressions that assess reporting bias, e.g., gender, presence of a spouse, the presence of at least one advisor outside Germany, mentioned funding, and mentioned type of employment. Women seem to engage in reporting bias less often, but this association is not statistically significant. Notably, economics dissertations and follow-up papers originating from economics dissertations show a positive association with reporting bias.

Finally, we also find that robustness checks and appendices are significantly positively associated with the share of statistical significant results when considering the whole dataset (1.4 percentage points difference). A possible explanation could be that authors find a statistically significant result in the main analysis and then run several robustness checks to solidify the robustness of the results.

In **Table 3.2**, we report results of the regressions for the share of statistically significant tests and statistical significance inside the 0.150 caliper as in **Table 3.1** but consider **only main tests** (i.e., further excluding obvious controls, correlations, robustness checks, and appendices). We also check the robustness of our results in **Table 3.1** by considering the significance levels of 10 percent and 1 percent.

Table 3.2—Main Regression including only Main Tests

	Dissertations			Dissertations/Papers		
	Number of Tests (Main Tests only) (1)	**Share Stat. Sig. (Main Tests only) (2)**	**Significant (Main Tests only) (3)**	**Number of Tests (Main Tests only) (4)**	**Share Stat. Sig. (Main Tests only) (5)**	**Significant (Main Tests only) (6)**
Panel A: 10%						
Graduate School	-18.341	0.003	-0.065			
	(24.376)	(0.012)	(0.022)			
Mandatory Supervision Agenda	71.702	-0.024	0.011			
	(45.067)	(0.013)	(0.034)			
Paper				-82.123	-0.003	-0.019
				(10.528)	(0.015)	(0.025)
Observations	**289**	**26,606**	**1,866**	**541**	**34,374**	**2,519**
Panel B: 5%						
Graduate School	-18.341	0.005	0.045			
	(24.376)	(0.014)	(0.034)			
Mandatory Supervision Agenda	71.702	-0.019	0.014			
	(45.067)	(0.012)	(0.044)			
Paper				-82.123	-0.001	-0.001
				(10.528)	(0.015)	(0.041)
Observations	**289**	**26,606**	**1,882**	**541**	**34,374**	**2,522**
Panel C: 1%						
Graduate School	-18.341	0.002	-0.050			
	(24.376)	(0.011)	(0.031)			
Mandatory Supervision Agenda	71.702	-0.023	-0.017			
	(45.067)	(0.011)	(0.047)			
Paper				-82.123	-0.000	0.022
				(10.528)	(0.016)	(0.037)
Observations	**289**	**26,606**	**1,309**	**541**	**34,374**	**1,739**
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes
Caliper	-	-	0.150	-	-	0.150

Note: We consider only the main tests (i.e., excluding intercepts, obvious controls, correlations, robustness checks, and appendices). Models 1-3 report regression results from dissertations only. Models 3-6 report regression results from dissertations and follow-up papers. In Models 1 and 4 we report OLS regression results. In Models 2-3 and 5-6, we report average marginal effects from logit regressions, with the outcome variable in Panels A, B, and C being an indicator variable for at least 10 percent, 5 percent, and 1 percent significance, respectively. In Models 3 and 6 consider only test statistics inside a 0.150 caliper around the corresponding *z*-value. Imprecise *z*-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Graduate School and Supervision Agenda dummies were kept fixed in Models 1-3 and were included in the Post-Double Lasso for Models 4-6 but are not displayed in the table. Year FE and Region FE were kept fixed. Standard errors are clustered at the university level in Models 1-3 and at the author level in Models 4-6.

Considering only the dissertation data, we do not find any systematic relationship between the presence of a graduate school and the share of statistically significant results at any of the three conventional significance levels. For the presence of mandatory supervision agendas, we find a systematic negative relationship with the share of statistical significant results. For the 10 percent level, we find a statistically significant negative relationship of -2.4

percentage points; for the 5 percent level, we also find a negative relationship (-1.9 percent points difference), but not statistically significant; and for the 1 percent level, we find a statistically significant negative relationship (-2.3 percentage points difference). Considering again only the dissertation data, it seems that the presence of graduate schools might be negatively associated with statistical significance inside the 0.150 caliper, but we find a negative statistically significant relationship for the 10 percent significance level (-6.5 percentage points difference), a positive but statistically insignificant relationship for the 5 percent significance level (4.5 percentage points difference), and a negative but statistically insignificant relationship for the 1 percent significance level (-5 percentage points difference). These results indicate that environments like graduate schools with course programs where students are taught research methodology might decrease questionable research practices like selective reporting. For the caliper regressions, we do not find any systematic relationship between mandatory supervision agendas and statistical significance inside the 0.150 caliper.

When considering the whole dataset, we still find a systematic negative relationship between the presence of a mandatory supervision agenda and the share of statistical significant results. For the 10 percent level, we find a statistically significant negative relationship (-3.3 percentage points difference); for the 5 percent, we find a statistically insignificant negative relationship (-2.8 percentage points); and for the 1 percent level, we find a statistically significant negative relationship (-3.7 percentage points difference). For the presence of graduate schools and for the caliper regressions, we do not find any systematic relationship of the two variables.

Finally, we observe the number of main test statistics is reduced from the dissertation to the paper stage, but we do not find that reporting bias emerges from the dissertation- to the paper stage, nor when we consider 10, 5, or 1 percent significance level. Specifically, for the 5 percent significance level, we do not see an increase in reporting bias either for the share of statistically significant main tests as the outcome (-0.1 percentage points difference) and for the indicator variable for statistical significance at 5 percent inside a 0.150 caliper (-0.1 percentage points difference). In short, it seems that the presence of mandatory supervision agendas is negatively associated with reporting bias. This result stays robust when considering only the main tests and when considering the 10 percent and 1 percent significance levels. For the presence of graduate schools, the results are less systematic but also indicate a negative relationship with reporting bias. Finally, the number of test statistics is reduced from the dissertation to the paper stage without the emergence of reporting bias. This result remains

stable when we consider only the main tests and also examine reporting bias at the significance levels of 10 or 1 percent.[11]

D. *Robustness Checks*

We report additional robustness checks in the Online Appendix where in **Table D4**, we consider only main tests and include the inverse of the number of tests per dissertation/follow-up papers as weights and find similar results. In **Online Appendix Table D5,** we report findings from regressions similar to the models in **Table 3.1** but without control variables (we still include year and region FE) and find similar results. Moreover, in **Online Appendix Table D6,** we consider only dissertations that never produced a follow-up paper and find a statistically significant negative relationship between the presence of a mandatory supervision agenda and the share of statistical significance at the 5 percent (-5,9 percent points difference). For the presence of a graduate school, we find a significant positive relationship with 5 percent significance inside the 0.150 caliper (7.6 percent points difference). In **Online Appendix Figure A5**, **Table C3**, and **Table D7** we focus only on the cumulative dissertation chapters and follow-up papers that clearly matched with each other. We still do not find any systematic indication of increased reporting bias from the dissertation to the paper stage. In **Figure A5**, no visible discontinuities can be seen around the common significance thresholds; the curves do not generally differ from each other. The binomial tests in **Table C3** indicate possible reporting bias in the follow-up paper subset for the 0.050 and 0.010 calipers. However, the regression results in **Table D7** do not indicate an increase in reporting bias from the dissertation to the paper stage. We also apply the tests of (Elliott, Kudrin, and Wüthrich 2022) as an additional robustness check and find the same results (see Online Appendix Table C6).

[11] Additionally, as specified in the pre-analysis plan, we run a beta regression for the share of statistical significance as the outcome in **Online Appendix Table D2** and find similar results as in our main tables. Lastly, as noted in Section I.B., we specified in the PAP an analysis regarding the introduction of the publication-based Handelsblatt researcher ranking introduced around 2007 for German-speaking countries in economics. However, due to a lack of variability in our data regarding the field and, therefore, difficulty in the interpretation of the results, we decided to report the results only in **Online Appendix Table D3**, where we run similar regressions as in our main regression tables but with some tweaks: 1) We do the analysis once for dissertations-only and for follow-up papers-only; 2) Instead of running logit regressions, we run linear probability models for Models 2-3 and 5-6 where we focus on the share of statistical significance and statistical significance inside the 0.150 caliper respectively. We find no systematic association between the introduction of the Handelsblatt ranking and reporting bias in economics.

III. Mechanisms

To understand why we see no reporting bias in dissertations and find no association of selective reporting with a reporting bias in papers resulting from these dissertations, we look at the publication process. First, we explore the presence of reporting bias in papers included in dissertations, but published after the defense and compare the extent of reporting bias between papers published before and after the defense. Second, we assess if reporting bias is associated with the impact factor of journals where papers from the dissertations were published.

A. *Publications before/after defense*

We divided our sample of follow-up papers into two groups: published before the PhD was defended and published after the PhD was defended to see the difference in reporting between them. We focus only on papers we could match to a cumulative dissertation chapter to make a clear comparison. More than 80 percent of the follow-up papers were published after the defense. On the density curve (see **Online Appendix Figure A4**), we see that papers published after the defense have a larger density around the 10 percent and 5 percent significance thresholds than those published before the defense. Overall, there is a shift between these groups, with the density for papers after the defense being higher for *z*-values below 2.58 and lower for *z*-values above 2.58 compared to papers published before the defense. In the next step, we apply binomial tests as in Section II.B, but now, for all follow-up papers originating from cumulative dissertation chapters and differentiating between those published before and after the defense.

The results in **Table 4** indicate reporting bias in the 0.050 and 0.010 calipers when considering all papers originating from cumulative dissertation chapters published after the defense. If we do the same analysis for the 10 percent and 1 percent levels (see **Online Appendix Tables C4 and C5**), we find that results are similar on the 1 percent level, while for the 10 percent level, we do not find systematic indications of reporting bias after the PhD defense. Lastly, we run regressions like Section II.C but again consider only the follow-up papers, and instead of the *paper* dummy, we include a dummy variable *after the defense*, taking value one if the paper was published after the defense and zero if it was published before the defense. Here, we do not find any indication of a relationship between the publication timing and reporting bias (see **Online Appendix Table D8**).

Table 4—Binomial Tests for the 5 percent significance level for follow-up papers that originated from a cumulative chapter overall and published before and after defense.

	All	**Before Defense**	**After Defense**
Caliper Size		0.150	
No. of Tests in Caliper	951	91	860
Under Caliper	469	50	419
Over Caliper	482	41	441
Binomial Probability	0.507	0.451	0.513
95% Confidence Interval	[0.476, 0.537]	[0.222, 0.567]	[0.484, 0.541]
Caliper Size		0.050	
No. of Tests in Caliper	315	34	281
Under Caliper	132	18	114
Over Caliper	183	16	167
Binomial Probability	0.581	0.471	0.594
95% Confidence Interval	[0.537, 0.625]	[0.300, 0.615]	[0.552, 0.640]
Caliper Size		0.010	
No. of Tests in Caliper	60	7	53
Under Caliper	25	4	21
Over Caliper	35	3	32
Binomial Probability	0.583	0.429	0.604
95% Confidence Interval	[0.486, 0.703]	[0.000, 0.600]	[0.500, 0.719]

Notes: In this table, we present the results of binomial proportion tests for test statistics for all follow-up papers that we could match to a cumulative dissertation chapter. Then, we disentangle it into follow-up papers published before and after the defense. A success is defined as a statistically significant observation at the 5 percent significance level. In the first panel, we use observations where ($1.81 < z < 2.16$); in the second panel, we use observations where ($1.91 < z < 2.01$); in the last panel, we use observations where ($1.95 < z < 1.97$). We then test if this proportion is statistically different from 0.5. We apply bootstrapping to consider non-independence between observations and report the 95 percent confidence intervals.

B. *Journal Impact Factors*

While previous studies mainly focused on top journals, we use a diverse and representative random sample and collect three different journal impact factors of each journal: 1) the overall impact factor collected from RePEc, 2) the 10-year impact factor collected from RePEc, and 3) The 5-year impact factor collected from the Web of Science database. This way, extend the approach of Brodeur, Cook, and Heyes (2020), who considered the 10-year impact factor by the RePEc.

The impact factors were retrieved by a research assistant from each website directly. One of the authors verified the data for possible uncertainties or errors to ensure sufficient data quality. We assigned a zero for journals with no available impact factor in **Table 5** Panel A , and as robustness check provided the same analysis only on journals with available impact factor **Table 5** Panel B. We consider only the follow-up papers with the share of statistically significant tests and the statistical significance inside the 0.150 calipers as the outcomes. For each of the three journal impact factors, we estimate the regressions with both outcomes,

respectively, with the journal impact factors as the explanatory variables and control variables selected by the machine learning algorithm.

The results in **Table 5** indicate a statistically significant positive association between the RePEc overall impact factor and the share of statistical significance (0.1 percentage point per point of impact factor). We also find positive associations between the RePEc overall and 10-year journal impact factors and statistical significance inside the 0.150 caliper (0.4 and 0.8 percentage points, respectively), which are both statistically significant.

Table 5—Impact Factor regression**.** Only follow-up papers considered.

	Follow-Up Papers					
	Share Stat. Sig. 5%	**Significant at 5%**	**Share Stat. Sig. 5%**	**Significant at 5%**	**Share Stat. Sig. 5%**	**Significant at 5%**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	Panel A: All journals					
RePEc (All)	0.001	0.004				
	(0.001)	(0.001)				
RePEc (10 Years)			0.001	0.008		
			(0.001)	(0.002)		
WoS (5 Years)					0.002	0.010
					(0.002)	(0.006)
Observations	21,430	1,347	21,430	1,347	21,430	1,347
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes
Caliper	-	[1.81, 2.11]	-	[1.81, 2.11]	-	[1.81, 2.11]
R^2 tjur	0.780	0.610	0.780	0.611	0.780	0.605
RMSE	0.233	0.312	0.233	0.312	0.233	0.315
F	81.109	11.517	82.481	11.573	82.037	12.130
	Panel B: Only journals with an impact factor					
RePEc (All)	0.001	0.004				
	(0.001)	(0.002)				
RePEc (10 Years)			0.001	0.008		
			(0.001)	(0.002)		
WoS (5 Years)					0.002	0.015
					(0.002)	(0.006)
Observations	19,523	1,221	19,523	1,221	19,952	1,221
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes
Caliper	-	[1.81, 2.11]	-	[1.81, 2.11]	-	[1.81, 2.11]
R^2 tjur	0.100	0.610	0.100	0.611	0.786	0.610
RMSE	0.225	0.312	0.226	0.312	0.230	0.313
F	65.516	10.961	65.906	11.016	73.114	11.328

Note: All Models report regression results from follow-up papers only. Panel A reports results including journals without an impact factor (zero in case no impact factor is available), while Panel B reports results excluding journals without an impact factor. Models 1, 3, and 5 report average marginal effects from logit regressions, with the outcome variable being an indicator variable for at least 10 percent significance. Models 2, 4, and 6 also report average marginal effects from logit regressions, but we consider only test statistics inside a 0.150 caliper around the 1.96 z-value, i.e., absolute *z*-values between 1.81 and 2.11. Imprecise *z*-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with Post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Year FE, Region FE, and the Impact Factor Variables were fixed throughout. Standard errors are clustered at the author level in all Models.

IV. Conclusion

While the citations of dissertations decreased over time (Larivière, Zuccala, and Archambault 2008), their follow-up papers are key drivers for disseminating knowledge created during doctoral studies (A.-M. Asanov et al. 2024). At the same time, a growing literature shows that just-significant results in job market papers published by PhD graduates are related to a higher chance of getting an academic position (Brodeur, Kattan, and Musumeci 2024). In this high-powered study, we utilize test statistics from empirical dissertations defended at German universities (327 dissertations and 68,990 test statistics) and from their follow-up papers (301 follow-up papers and 25,861 test statistics), which, to our knowledge, is the largest manually collected dataset for analyzing reporting bias.

We do not find any indications for reporting bias through visual inspection of Gaussian density curves. Binomial tests also do not show systematic indications for reporting bias inside narrow calipers when considering the commonly inspected 5 percent and 10 percent thresholds. We observe selective reporting, as the number of tests sharply decreases from the dissertation to the paper stage. However, our regression analysis does not indicate an increase in reporting bias in published papers compared to dissertations. One could argue that dissertations here serve as "populated pre-analysis plans" (Banerjee et al., 2020), where the results of all estimations are published, which are then filtered out during the publication process without an apparent substantial preference for significant results. In addition, regression results show that the presence of a mandatory supervision agenda is negatively associated with reporting bias.

We explore possible mechanisms for the unexpected absence of reporting bias despite the presence of selective reporting. Thus, we study when (before/after defense) and where (type of journal) publication bias could emerge during the publication process. First, given that collegial and conservative institutional environments during the PhD phase seem to be responsible for the apparent absence of reporting bias in dissertations, we assess if papers published after the defense are at risk of reporting bias. While we see that papers published after the defense are susceptible to reporting bias, it still does not seem to lead to an increase in reporting bias in published papers compared to dissertations.

Second, "unbiased" selective reporting can mask self-selection on statistical significance into journals with higher impact factor, especially given that papers resulting from the dissertations in our sample are published in a wide variety of journals - from the MDPI type

of journals to the American Economic Review. Indeed, we find that higher journal impact factors are positively associated with reporting just-significant results at the 5 percent level and with the overall share of statistical significant results. A one-point impact factor (in 10 years of RePec) is associated with about one percentage point increase in the share of statistically significant results. Thus, selective reporting bias on positive statistical significance seems to be less present in less competitive journals compared to the more competitive ones (the common focus of previous studies), reconciling the previous literature and our finding of "unbiased" selective reporting in papers resulting from a representative sample of dissertations.

While our study focuses on individuals who completed their PhD at German universities, we still consider our results generalizable. Follow-up papers from our representative sample of dissertations were published in various journals located across the world. Moreover, we also considered non-top journals, which was possible due to the random sampling of the dissertations. Finally, we focus on the earliest stage of the research career, irrespective of whether the researcher enters the academic job market. Individuals are prone to engage in reporting bias if they decide to pursue an academic career (Brodeur, Kattan, and Musumeci 2024). Our results are surprising in that we generally do not find any systematic indication of reporting bias, which is uncommon in the literature.

Possibilities for further research include studies shedding light on supervisor characteristics since the student-supervisor relationship is one of the integral parts of PhD studies. Regarding limitations, the data in our study did not allow us to identify job market papers, which might be interesting for further studies. Lastly, we cannot claim causality with our analysis.

REFERENCES

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?*." *The Quarterly Journal of Economics* 138 (1): 1–35. https://doi.org/10.1093/qje/qjac038.

Asanov, Anastasiya-Mariya, Igor Asanov, Guido Buenstorf, Valon Kadriu, and Pia Schoch. 2024. "Patterns of Dissertation Dissemination: Publication-Based Outcomes of Doctoral Theses in the Social Sciences." *Scientometrics*, February. https://doi.org/10.1007/s11192-024-04952-1.

Asanov, Igor, Christoph Bühren, and Panagiota Zacharodimou. 2020. "The Power of Experiments: How Big Is Your n?" Working Paper 32–2020. MAGKS Joint Discussion Paper Series in Economics. https://www.econstor.eu/handle/10419/234837.

Askarov, Zohid, Anthony Doucouliagos, Hristos Doucouliagos, and T D Stanley. 2023. "The Significance of Data-Sharing Policy." *Journal of the European Economic Association* 21 (3): 1191–1226. https://doi.org/10.1093/jeea/jvac053.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics. No. w26993. National Bureau of Economic Research, 2020.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls†." *The Review of Economic Studies* 81 (2): 608–50. https://doi.org/10.1093/restud/rdt044.

Blanco-Perez, Cristina, and Abel Brodeur. 2020. "Publication Bias and Editorial Statement on Negative Findings." *The Economic Journal* 130 (629): 1226–47. https://doi.org/10.1093/ej/ueaa011.

Brodeur, Abel, Scott Carrell, David Figlio, and Lester Lusher. 2023. "Unpacking P-Hacking and Publication Bias." *American Economic Review* 113 (11): 2974–3002. https://doi.org/10.1257/aer.20210795.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60. https://doi.org/10.1257/aer.20190687.

Brodeur, Abel, Nikolai M Cook, Jonathan S Hartley, and Anthony Heyes. 2024. "Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement," no. 101.

Brodeur, Abel, Nikolai Cook, and Carina Neisser. 2024. “P-Hacking, Data Type and Data-Sharing Policy.” *The Economic Journal* 134 (659): 985–1018. https://doi.org/10.1093/ej/uead104.

Brodeur, Abel, Lamis Kattan, and Marco Musumeci. 2024. “Job Market Stars.” Working Paper 141. I4R Discussion Paper Series. https://www.econstor.eu/handle/10419/301430.

Brodeur, Abel, David Valenta, Alexandru Marcoci, Juan P. Aparicio, Derek Mikola, Bruno Barbarioli, Rohan Alexander, et al. 2025. “Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science.” Working Paper 195. I4R Discussion Paper Series. https://www.econstor.eu/handle/10419/308508.

Bruns, Stephan B., Teshome K. Deressa, T. D. Stanley, Chris Doucouliagos, and John P. A. Ioannidis. 2024. “Estimating the Extent of Selective Reporting: An Application to Economics.” *Research Synthesis Methods* n/a (n/a). https://doi.org/10.1002/jrsm.1711.

Bruns, Stephan B., and Martin Kalthaus. 2020. “Flexibility in the Selection of Patent Counts: Implications for p-Hacking and Evidence-Based Policymaking.” *Research Policy* 49 (1): 103877. https://doi.org/10.1016/j.respol.2019.103877.

Chopra, Felix, Ingar Haaland, Christopher Roth, and Andreas Stegmann. 2023. “The Null Result Penalty.” *The Economic Journal*, August, uead060. https://doi.org/10.1093/ej/uead060.

Cilliers, Jacobus, Nour Elashmawy, and David McKenzie. 2024. “Using Post-Double Selection Lasso in Field Experiments.” Washington, DC: World Bank. https://doi.org/10.1596/1813-9450-10931.

Doucouliagos, Hristos, Thomas Hinz, and Katarina Zigova. 2022. “Bias and Careers: Evidence from the Aid Effectiveness Literature.” *European Journal of Political Economy* 71 (January):102056. https://doi.org/10.1016/j.ejpoleco.2021.102056.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wüthrich. 2022. “Detecting P-Hacking.” *Econometrica* 90 (2): 887–906. https://doi.org/10.3982/ECTA18583.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. “Publication Bias in the Social Sciences: Unlocking the File Drawer.” *Science* 345 (6203): 1502–5. https://doi.org/10.1126/science.1255484.

Gerber, Alan, and Neil Malhotra. 2008a. “Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals.” *Quarterly Journal of Political Science* 3 (3): 313–26. https://doi.org/10.1561/100.00008024.

———. 2008b. “Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?” *Sociological Methods & Research* 37 (1): 3–30. https://doi.org/10.1177/0049124108318973.

Hüther, Otto, and Georg Krücken. 2018. *Higher Education in Germany—Recent Developments in an International Perspective*. Vol. 49. Higher Education Dynamics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-61479-3.

Imbens, Guido W. 2021. “Statistical Significance,p-Values, and the Reporting of Uncertainty.” *Journal of Economic Perspectives* 35 (3): 157–74. https://doi.org/10.1257/jep.35.3.157.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. “The Power of Bias in Economics Research.” *The Economic Journal* 127 (605): F236–65. https://doi.org/10.1111/ecoj.12461.

Kranz, Sebastian, and Peter Pütz. 2022. “Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics: Comment.” *American Economic Review* 112 (9): 3124–36. https://doi.org/10.1257/aer.20210121.

Larivière, Vincent, Alesia Zuccala, and Éric Archambault. 2008. “The Declining Scientific Impact of Theses: Implications for Electronic Thesis and Dissertation Repositories and Graduate Studies.” *Scientometrics* 74 (1): 109–21. https://doi.org/10.1007/s11192-008-0106-3.

Roebken, H. 2007. “Postgraduate Studies in Germany - How Much Structure Is Not Enough?” *South African Journal of Higher Education* 21 (8): 1054–66.

Vivalt, Eva. 2019. “Specification Searching and Significance Inflation Across Time, Methods and Disciplines.” *Oxford Bulletin of Economics and Statistics* 81 (4): 797–816. https://doi.org/10.1111/obes.12289.

Unbiased Selective Reporting? The Pursuit of Publishable Results by Early-Stage Researchers (Online Appendix)

Authors

Anastasiya-Mariya Asanov (ORCID: 0000-0003-3080-4213; noha@incher.uni-kassel.de) [1]

Igor Asanov (ORCID: 0000-0002-8091-4130; igor.asanov@uni-kassel.de) [1,*]

Guido Buenstorf (ORCID: 0000-0002-2957-5532; buenstorf@uni-kassel.de) [1]

Valon Kadriu (ORCID: 0009-0006-1101-5349; kadriu@uni-kassel.de) [1]

Pia Schoch (ORCID: 0009-0006-9471-4590; p.schoch@uni-kassel.de) [1]

[1] University of Kassel, INCHER and Institute of Economics, Kassel, Germany

[*] Corresponding author (e-mail: igor.asanov@uni-kassel.de)

Online Appendix 1: Additional Figures

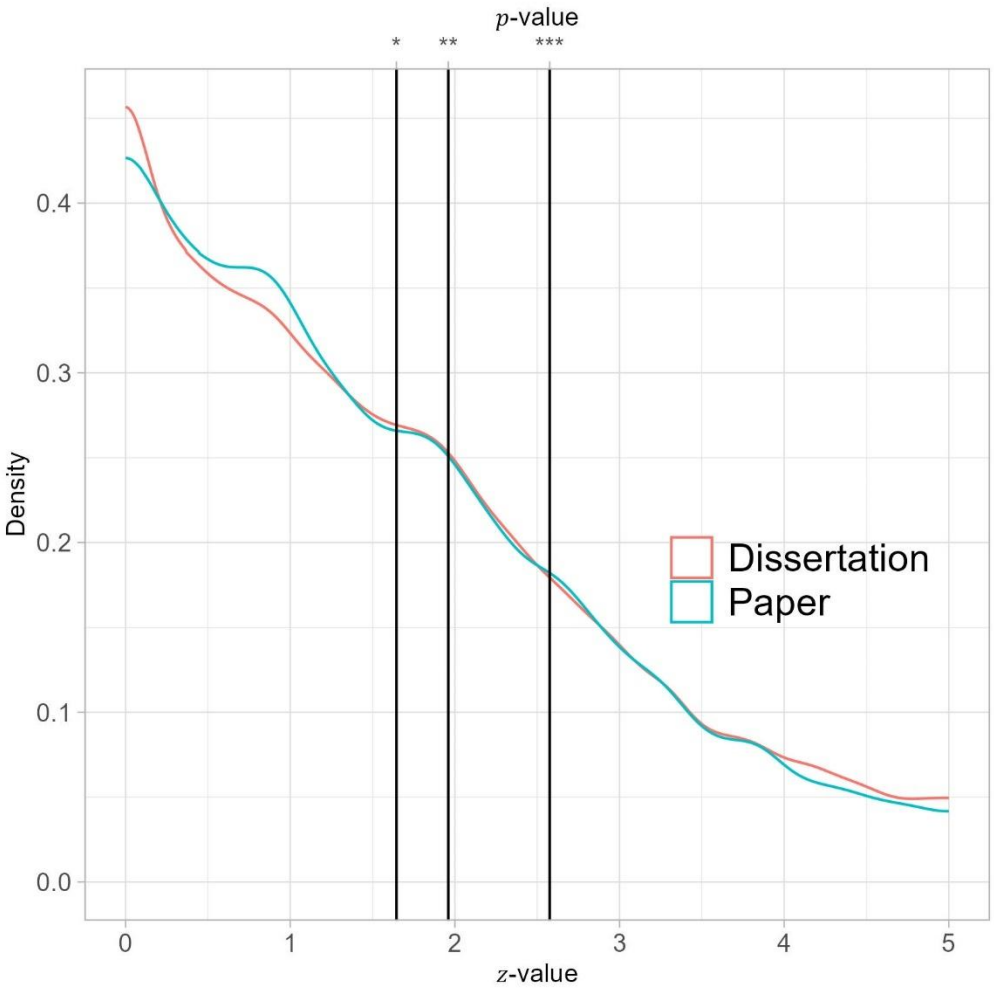

Figure A1. Gaussian density lines of a histogram with z-values of test statistics from follow-up papers of dissertations and the dissertations themselves considering only $z \in [0, 5]$. Black vertical lines depict z-values corresponding to the conventional significance thresholds for 10 percent, 5 percent, and 1 percent. Imprecisely reported z-values removed following (Kranz and Pütz 2022).

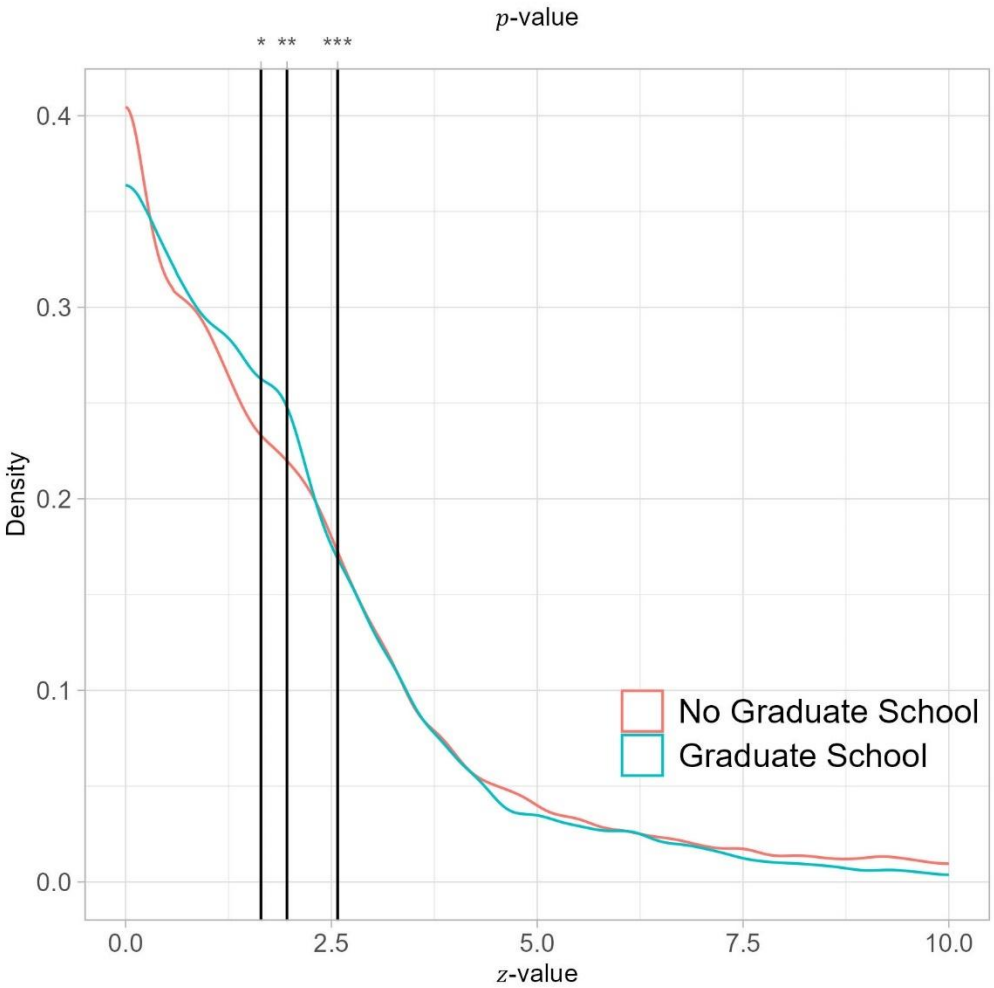

Figure A2. Gaussian density lines of z-values of test statistics from follow-up studies of dissertations and the dissertations themselves considering the presence/absence of graduate schools during time of PhD completion and considering only $z \in [0, 10]$. Black vertical lines depict z-values corresponding to the conventional significance thresholds for 10 percent, 5 percent, and 1 percent. Imprecisely reported z-values removed following (Kranz and Pütz 2022).

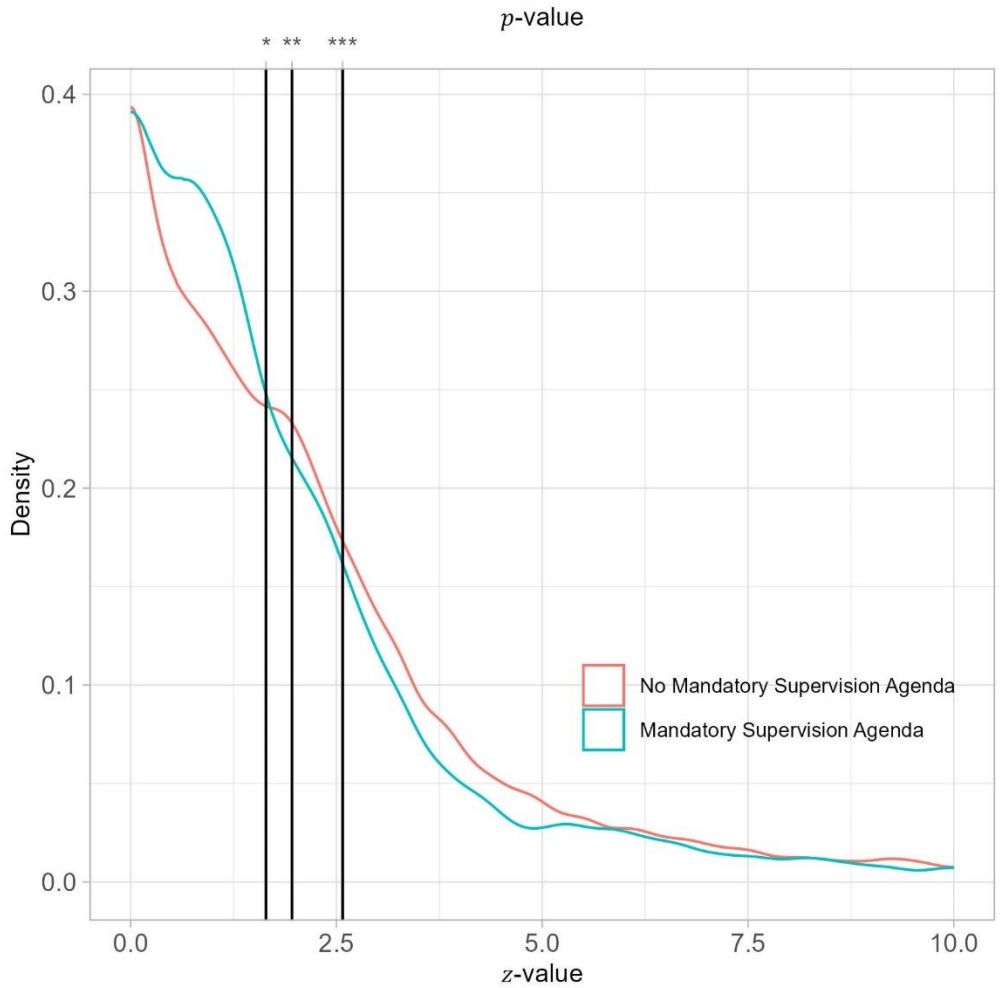

Figure A3. Gaussian density lines of z-values of test statistics from follow-up papers of dissertations and the dissertations themselves considering the presence/absence of mandatory supervision agendas during time of PhD completion and considering only $z \in [0, 10]$. Black vertical lines depict z-values corresponding to the conventional significance thresholds for 10 percent, 5 percent, and 1 percent. Imprecisely reported *z*-values removed following (Kranz and Pütz 2022).

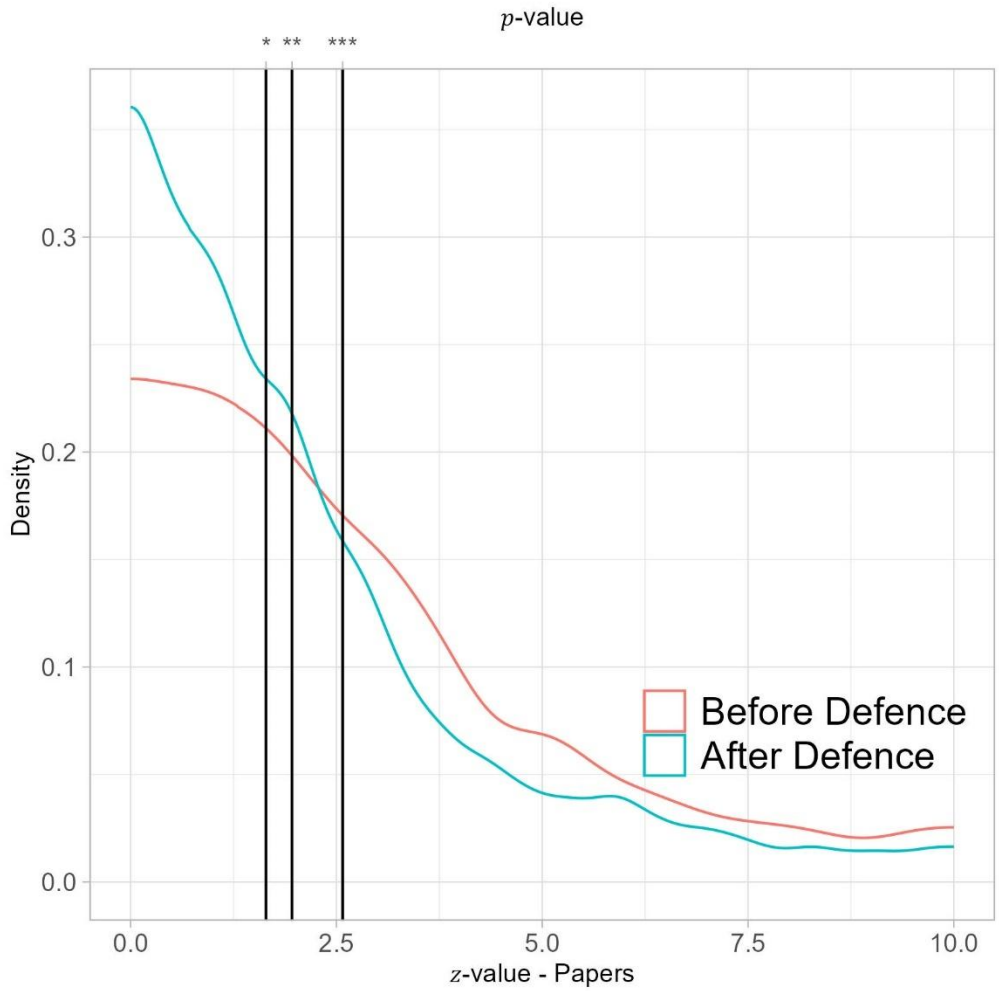

Figure A4. Gaussian density lines of a histogram with *z*-values of test statistics from follow-up papers that were published before the defense (red solid line) and after the defense (blue solid line). Black vertical lines depict *z*-values corresponding to the conventional significance thresholds for 10 percent, 5 percent, and 1 percent. Imprecisely reported *z*-values removed following (Kranz and Pütz 2022).

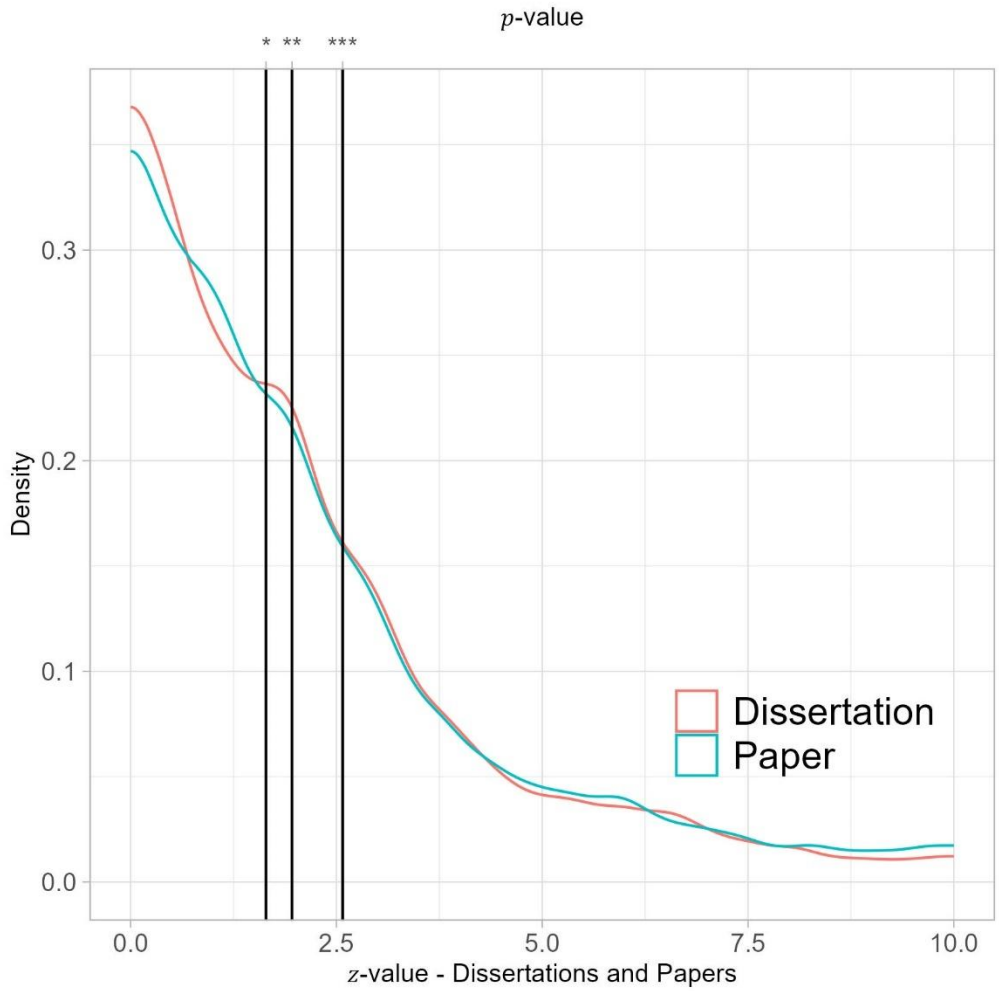

Figure A5. Gaussian density lines of a histogram with *z*-values of test statistics from follow-up papers of dissertations and the dissertations themselves. Here, we only consider 1) cumulative dissertation chapters that later got published as a follow-up paper and 2) the follow-up papers that originated from those chapters. Black vertical lines depict *z*-values corresponding to the conventional significance thresholds for 10 percent, 5 percent, and 1 percent. Imprecisely reported *z*-values removed following (Kranz and Pütz 2022).

Online Appendix 2: Additional Tables

A List of Control Variables

Table A—List of Control Variables

Variable	**Description**
University Level Variables	
University Type	Categorical variable with categories: university, technical university, free university.
Old University	Binary variable equal to 1 if the university was founded before 1945.
Number of Students	Number of students in the year when the dissertation was submitted
Number of Professors	Number of professors in the year when the dissertation was submitted *not included because more that 20% missing values (30% missing)
University Region	Categorical with 16 categories for the regions of Germany
City Population	City population in the year when the dissertation was submitted
Dissertation and Author Level Variables	
Field	Categorical variable for the field in which the dissertation was written: Economics Sociology Political Science *used economics dummy instead as moajority of dissertations are in economics.
English	A binary variable equal to 1 if the dissertation was written in English *not included because

	not enough observations
Year of the Dissertation	A categorical variable for the year the dissertation was published: 2004, 2005, 2006, 2012, 2013, 2014 *added due to the data collection procedure
Number of Pages	Number of pages in the dissertation
Number of Chapters	Number of chapters in the dissertation
Number of Advisors	Number of advisors mentioned
Advisor from Abroad	Binary variable equal to 1 if at least one advisor works outside o Germany
Principal Component based on keywords	First principal component calculated based on keywords assigned to the dissertation in the German National Library **not included because not enough observations
Place of Birth outside of Germany	Binary variable equal to 1 if the author was born outside of Germany **not included because not enough observations
Female	Binary variable equal to 1 if the author is female
Spouse	Binary variable equal to 1 if a spouse is mentioned in acknowledgements
Children	Binary variable equal to 1 if a children are mentioned in acknowledgements * not available
Age	Calculated based on the date of birth on the front page or CV attached to the dissertation *removed because not available
International Education	Binary variable equal to 1 if the author received any education outside of Gemany

Affiliation with the Max-Planck institute	Binary variable equal to 1 if affiliation to the Max-Planck Institute is mentioned in the acknowledgements *not included because not enough observations
Mentioned funding	Binary variable equal to 1 if receiving funding from the university or government is mentioned in the acknowledgements
Mentioned employment	Binary variable equal to 1 if employment at the university or institute is mentioned in the acknowledgements
Eye Catchers	Binary: equal to 1 if stars or other eye catchers are used to signal statistical significance
Formal Model	Binary: equal to 1 if a formal model is used in the paper
Own Data	Binary: equal to 1 if data was collected by the authors, e.g. surveys and interviews
External Data	Binary: equal to 1 if external data sources were used
Cross Section	Binary: equal to 1 if cross section data
Time Series	Binary: equal to 1 if time series data
Panel	Binary equal to 1 if panel data
Lab Experiment	Binary: equal to 1 if Lab Experiment
Field Experiemnt	Binary: equal to 1 if Field Experiment
Quasi-Experiment	Binary: equal to 1 if Quasi-Experiment

Observational	Binary: equal to 1 if observational
IV	Binary: equal to 1 if IV
RDD	Binary: equal to 1 if RDD
DID	Binary: equal to 1 if DID
RCT	Binary: equal to 1 if RCT
Number of Observations	Numeric: Number of observations per regression model
S20	Binary: equal to 1 if stars in Notes for 20% significance are used *not included because not enough observations
S15	Binary: equal to 1 if stars in Notes for 15% significance are used
S10	Binary: equal to 1 if stars in Notes for 10% significance are used
S05	Binary: equal to 1 if stars in Notes for 5% significance are used
S01	Binary: equal to 1 if stars in Notes for 1% significance are used
S001	Binary: equal to 1 if stars in Notes for 0.1% significance are used
Robustness/Extension	Binary: equal to 1 if model is declared to be robustness check
Two-sided	Binary: Equal to 1 if it is a two-sided test

Reported Significance	Reported Significance by means of eye-catcher

Notes: This set of variables was included in the pre-analysis plan. We specified that a variable would be included in the analysis if it is available in more than 80% of observations.

B Journal Locations

Table B—Journal location of journal articles

	Follow-Up Papers		
	%	No. of papers	No.of Teststats
Journal Location			
Europe	**75%**	**227**	**21,055**
United Kingdom	44%	99	8,149
Netherlands	31%	71	6,854
Germany	19%	44	4,317
Switzerland	4%	9	1,178
Czech Republic	1%	2	441
France	0.5%	1	97
Italy	0.5%	1	19
North America	**21%**	**64**	**4,433**
United States	97%	62	4,270
Canada	3%	2	163
South America	**2%**	**5**	**257**
Colombia	40%	2	146
Chile	40%	2	62
Bolivia	20%	1	49
Asia	**1%**	**2**	**61**
China	50%	1	58
Singapore	50%	1	3
Africa	**1%**	**3**	**55**
Ethiopia	67%	2	41
Nigeria	33%	1	14

C Binomial Tests

Table C1—Binomial Tests for the 10 percent significance level

	All	Dissertation	Paper
Caliper Size		0.150	
No. of Tests in Caliper	6,344	4,624	1,720
Under Caliper	3,207	2,363	844
Over Caliper	3,137	2,261	876
Binomial Probability	0.494	0.489	0.509
95% Confidence Interval	[0.482, 0.507]	[0.476, 0.502]	[0.484, 0.534]
Caliper Size		0.050	
No. of Tests in Caliper	2,044	1,495	549
Under Caliper	1,097	797	300
Over Caliper	947	698	249
Binomial Probability	0.463	0.467	0.454
95% Confidence Interval	[0.443, 0.484]	[0.447, 0.487]	[0.413, 0.502]
Caliper Size		0.010	
No. of Tests in Caliper	427	298	129
Under Caliper	245	168	77
Over Caliper	182	130	52
Binomial Probability	0.426	0.436	0.403
95% Confidence Interval	[0.380, 0.473]	[0.385, 0.487]	[0.317, 0.522]

Notes: In this table, we present the results of binomial proportion tests for test statistics for our overall dataset and for each test statistic from dissertations and their follow-up papers where a success is defined as a statistically significant observation at the 10 percent significance level. In the first panel, we use observations where ($1.49 < z < 1.79$); in the second panel, we use observations where ($1.59 < z < 1.69$); in the last panel we use observations where ($1.63 < z < 1.65$). We then test if this proportion is statistically different from 0.5. The associated 95 percent confidence intervals are reported. We apply bootstrapping to consider non-independence between observations.

Table C2—Binomial Tests for the 1 percent significance level

	All	**Dissertation**	**Paper**
Caliper Size		0.150	
No. of Tests in Caliper	4,226	3,054	1,172
Under Caliper	1,987	1,476	511
Over Caliper	2,239	1,578	661
Binomial Probability	**0.530**	0.517	**0.564**
95% Confidence Interval	**[0.513, 0.547]**	[0.499, 0.535]	**[0.533, 0.593]**
Caliper Size		0.050	
No. of Tests in Caliper	1,490	1,047	443
Under Caliper	582	427	155
Over Caliper	908	620	288
Binomial Probability	**0.609**	**0.592**	**0.650**
95% Confidence Interval	**[0.579, 0.638]**	**[0.564, 0.620]**	**[0.593, 0.699]**
Caliper Size		0.010	
No. of Tests in Caliper	431	278	153
Under Caliper	114	86	28
Over Caliper	317	192	125
Binomial Probability	**0.735**	**0.691**	**0.817**
95% Confidence Interval	**[0.674, 0.786]**	**[0.610, 0.750]**	**[0.731, 0.871]**

Notes: In this table, we present the results of binomial proportion tests for test statistics for our overall dataset and for each test statistic from dissertations and their follow-up papers where a success is defined as a statistically significant observation at the 1 percent significance level. In the first panel, we use observations where ($2.43 < z < 2.73$); in the second panel, we use observations where ($2.53 < z < 2.63$); in the last panel, we use observations where ($2.57 < z < 2.59$). We then test if this proportion is statistically different from 0.5. The associated 95 percent confidence intervals are reported. We apply bootstrapping to consider non-independence between observations.

Table C3—Binomial Tests for the 5 percent significance level for overall matched data, the cumulative dissertation chapters that got later published, and the corresponding follow-up papers

	All	**Dissertation**	**Paper**
Caliper Size		0.150	
No. of Tests in Caliper	2,081	1,130	951
Under Caliper	1,047	578	469
Over Caliper	1,034	552	482
Binomial Probability	0.497	0.488	0.507
95% Confidence Interval	[0.473, 0.521]	[0.461, 0.515]	[0.476, 0.537]
Caliper Size		0.050	
No. of Tests in Caliper	699	384	315
Under Caliper	313	181	132
Over Caliper	386	203	183
Binomial Probability	**0.552**	0.529	**0.581**
95% Confidence Interval	**[0.520, 0.581]**	[0.493, 0.561]	**[0.537, 0.625]**
Caliper Size		0.010	
No. of Tests in Caliper	134	74	60
Under Caliper	59	34	25
Over Caliper	75	40	35
Binomial Probability	0.560	0.541	0.583
95% Confidence Interval	[0.493, 0.635]	[0.460, 0.615]	[0.486, 0.703]

Notes: In this table, we present the results of binomial proportion tests for test statistics for our overall dataset and for each test statistic from dissertations and their follow-up papers where a success is defined as a statistically significant observation at the 5 percent significance level. In the first panel, we use observations where ($1.81 < z < 2.11$); in the second panel, we use observations where ($1.91 < z < 2.01$); in the last panel, we use observations where ($1.95 < z < 1.97$). We then test if this proportion is statistically different from 0.5. We apply bootstrapping to consider non-independence between observations and report the 95 percent confidence intervals.

Table C4—Binomial Tests for the 10 percent significance level for follow-up papers that originated from a cumulative chapter, the follow-up papers that got published before the dissertation defense, and the follow-up papers that got published after the dissertation defense

	All	**Before Defense**	**After Defense**
Caliper Size		0.150	
No. of Tests in Caliper	969	102	867
Under Caliper	455	53	402
Over Caliper	514	49	465
Binomial Probability	**0.530**	0.480	**0.536**
95% Confidence Interval	**[0.505, 0.562]**	[0.390, 0.590]	**[0.507, 0.570]**
Caliper Size		0.050	
No. of Tests in Caliper	306	29	277
Under Caliper	154	16	138
Over Caliper	152	13	139
Binomial Probability	0.497	0.448	0.502
95% Confidence Interval	[0.449, 0.550]	[0.300, 0.667]	[0.455, 0.563]
Caliper Size		0.010	
No. of Tests in Caliper	70	8	62
Under Caliper	36	5	31
Over Caliper	34	3	31
Binomial Probability	0.486	0.375	0.500
95% Confidence Interval	[0.381, 0.609]	[0.163, 0.667]	[0.382, 0.636]

Notes: In this table, we present the results of binomial proportion tests for test statistics for all follow-up papers that we could match to a cumulative dissertation chapter. Then, we disentangle it into follow-up papers published before and after the defense. A success is defined as a statistically significant observation at the 10 percent significance level. In the first panel, we use observations where ($1.49 < z < 1.79$); in the second panel, we use observations where ($1.59 < z < 1.69$); in the last panel, we use observations where ($1.63 < z < 1.65$). We then test if this proportion is statistically different from 0.5. We apply bootstrapping to consider non-independence between observations and report the 95 percent confidence intervals.

Table C5—Binomial Tests for the 1 percent significance level for follow-up papers that originated from a cumulative chapter, the follow-up papers that got published before the dissertation defense, and the follow-up papers that got published after the dissertation defense

	All	**Before Defense**	**After Defense**
Caliper Size		0.150	
No. of Tests in Caliper	666	69	597
Under Caliper	308	32	276
Over Caliper	358	37	321
Binomial Probability	0.538	0.536	0.538
95% Confidence Interval	[0.499, 0.574]	[0.432, 0.615]	[0.497, 0.576]
Caliper Size		0.050	
No. of Tests in Caliper	228	30	198
Under Caliper	89	12	77
Over Caliper	139	18	121
Binomial Probability	**0.610**	0.600	**0.611**
95% Confidence Interval	**[0.543, 0.674]**	[0.444, 0.765]	**[0.548, 0.676]**
Caliper Size		0.010	
No. of Tests in Caliper	72		70
Under Caliper	22	0	22
Over Caliper	50	2	48
Binomial Probability	**0.694**	**1.000**	**0.686**
95% Confidence Interval	**[0.556, 0.776]**	**[1.000, 1.000]**	**[0.545, 0.769]**

Notes: In this table, we present the results of binomial proportion tests for test statistics for all follow-up papers that we could match to a cumulative dissertation chapter. Then, we disentangle it into follow-up papers published before and after the defense. A success is defined as a statistically significant observation at the 1 percent significance level. In the first panel, we use observations where ($2.43 < z < 2.73$); in the second panel, we use observations where ($2.53 < z < 2.63$); in the last panel, we use observations where ($2.57 < z < 2.59$). We then test if this proportion is statistically different from 0.5. We apply bootstrapping to consider non-independence between observations and report the 95 percent confidence intervals.

Table C6—Elliott, Kudrin, and Wüthirch's (2022) Tests

Threshold:	1% Significance		5% Significance		10% Significance				
Sample:	Bin.	Discont.	Bin.	Discont.	Bin.	Discont.	CS1	CS2B	LCM
All	1.000	0.465	0.539	0.465	0.999	0.465	0.000	0.000	0.930
Dissertations	1.000	0.245	0.397	0.245	0.999	0.245	0.004	0.000	1.000
Follow-Up Paper	1.000	0.567	0.767	0.567	0.862	0.567	0.016	0.025	0.968

D Regressions

Table D1—Main Regression (controls displayed)

	Dissertations			Dissertations/Papers		
	Number of Tests (1)	**Share Stat. Sig. 5% (2)**	**Significant at 5% (3)**	**Number of Tests (4)**	**Share Stat. Sig. 5% (5)**	**Significant at 5% (6)**
Variables of Interest						
Graduate School	-26.707	-0.001	0.007		-0.018	-0.006
	(39.225)	(0.009)	(0.028)		(0.010)	(0.022)
Mandatory Supervision Agenda	104.511	-0.003	-0.081		-0.006	
	(71.118)	(0.013)	(0.032)		(0.014)	
Paper				-150.037	-0.010	-0.034
				(19.829)	(0.012)	(0.029)
PDL Controls						
Share 5% Eye Catcher	109.051			77.671		
	(35.949)			(22.505)		
Share 1% Eye Catcher	103.606			40.721		
	(38.143)			(18.736)		
15% Eye Catcher		-0.060			-0.041	
		(0.024)			(0.020)	
10% Eye Catcher		-0.089			-0.017	
		(0.019)			(0.015)	
5% Eye Catcher					-0.105	
					(0.048)	
1% Eye Catcher					-0.060	
					(0.042)	
0.1% Eye Catcher		-0.052			-0.003	
		(0.026)			(0.023)	
Eye Catchers (general)		-0.050				
		(0.025)				
Share RDD	842.995			-29.485		
	(156.913)			(30.852)		
Share DID	173.416					
	(108.142)					
DID		-0.015			-0.008	
		(0.012)			(0.019)	
IV		0.002	0.016		0.013	
		(0.022)	(0.024)		(0.021)	
RDD		-0.013	-0.076		-0.016	
		(0.017)	(0.044)		(0.025)	
RCT					-0.012	
					(0.027)	
Economics		0.009	0.021		0.009	0.014
		(0.017)	(0.032)		(0.018)	(0.028)
English		0.013			0.014	-0.041
		(0.010)			(0.016)	(0.030)
Cumulative Dissertation		0.011		59.603	-0.004	0.024
		(0.016)		(23.808)	(0.013)	(0.025)
Share Formal Model				14.694		
				(14.698)		
Formal Model		0.047			0.006	0.015
		(0.037)			(0.013)	(0.034)
Cross-Section		-0.017			-0.024	

	Dissertations			Dissertations/Papers		
	Number of Tests (1)	**Share Stat. Sig. 5% (2)**	**Significant at 5% (3)**	**Number of Tests (4)**	**Share Stat. Sig. 5% (5)**	**Significant at 5% (6)**
		(0.015)			(0.013)	
Panel		-0.015	-0.001		-0.015	
		(0.011)	(0.024)		(0.013)	
Share Time Series				1.869		
				(33.896)		
Time Series		0.001	0.005		-0.006	-0.009
		(0.011)	(0.032)		(0.013)	(0.027)
External Data		-0.055	-0.052		-0.003	
		(0.042)	(0.030)		(0.017)	
Own Data		-0.070			-0.026	
		(0.042)			(0.019)	
Observational		-0.000			-0.009	-0.014
		(0.020)			(0.016)	(0.021)
Field Experiment		-0.029				
		(0.024)				
Lab Experiment		0.061	-0.013			
		(0.038)	(0.048)			
Quasi Experiment					-0.000	
					(0.016)	
Female		-0.006			-0.007	
		(0.012)			(0.011)	
Spouse		0.008	0.022		-0.006	
		(0.007)	(0.021)		(0.010)	
At least one Advisor from outside Germany		0.016	-0.003		0.018	0.006
		(0.013)	(0.039)		(0.013)	(0.023)
Mentioned Funding		-0.017	-0.009		0.006	
		(0.011)	(0.022)		(0.009)	
Mentioned Type of Employment		-0.002			-0.005	0.024
		(0.010)			(0.009)	(0.021)
Number of Chapters		0.006			0.000	-0.000
		(0.002)			(0.000)	(0.000)
Number of Pages					-0.000	
					(0.000)	
Old University		0.005			-0.008	0.013
		(0.009)			(0.009)	(0.022)
Obvious Control		-0.001	-0.016		0.005	
		(0.007)	(0.028)		(0.006)	
Robustness Check		0.013	-0.010		0.014	
		(0.010)	(0.023)		(0.007)	
Two-sided Test		-0.281			-0.095*	-0.523
		(0.058)			(0.051)	(0.008)
(Intercept)	127.051			124.614		
	(67.279)			(49.927)		
Observations	327	54565	3433	626	72390	5263
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
University Controls	No	Yes	Yes	No	Yes	Yes
Reported Sig. as Control	No	Yes	No	No	Yes	No
Caliper	-	-	[1.81, 2.11]	-	-	[1.81, 2.11]
R^2	0.129	-	-	0.140	-	-

	Dissertations			Dissertations/Papers		
	Number of Tests (1)	**Share Stat. Sig. 5% (2)**	**Significant at 5% (3)**	**Number of Tests (4)**	**Share Stat. Sig. 5% (5)**	**Significant at 5% (6)**
Adj. R^2	0.050	-	-	0.098	-	-
R^2 tjur	-	0.817	0.011	-	0.803	0.020
RMSE	-	0.212	0.497	-	0.221	0.495
F	-	149.65	1.079	-	232.88	0.997

Models 1-3 report regression results from dissertations only. Models 4-6 report regression results from dissertations and follow-up studies. In Models 1 and 4, we apply OLS regression, with the outcome variable being the count of test statistics per dissertation. Models 2-3 and 5-6 report average marginal effects from logit regressions, with the outcome variable in Models 2 and 5 being an indicator variable for at least 5 percent significance and an indicator variable for at least 5 percent significance in Models 3 and 6. Models 3 and 6 consider only test statistics inside a 0.150 caliper around the 1.96 z-value, i.e. absolute z-values between 1.81 and 2.11. Imprecise z-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with Post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Year FE and Region FE were kept fixed. Standard errors are clustered at the university level in Models 1-3 and at the author level in Models 4-6.

Table D2—Beta regression for dissertations only and for the overall dataset

	Dissertations	**Diss/Follow-Up Papers**
	Share stat. Sig. 5%	**Share stat. Sig. 5%**
	(1)	**(2)**
Graduate School	0.032	
	(0.098)	
Mandatory Supervision Agenda	-0.037	
	(0.148)	
Paper		0.006
		(0.089)
Observations	298	569
Year FE	Yes	Yes
Region FE	Yes	Yes
Reported Sig. as Control	Yes	Yes
Other Controls	Yes	Yes
Pseudo R^2	0.268	0.251
Log Likelihood	105.126	170.841

Model 1 reports regression results from dissertations only. Model 2 reports regression results from dissertations and follow-up papers. In both Models, we apply beta regression where the outcome variable is continuous for the share of statistically significant test statistics per dissertation or paper at a 5 percent level. Imprecise *z*-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with Post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Year FE and Region FE were kept fixed. Standard errors are clustered at the university level in Model 1 and the author level in Model 2.

Table D3—Handelsblatt Ranking Regression

	Dissertations			Follow-Up Papers		
	Number of Tests	**Share Stat. Sig. 5%**	**Significant at 5%**	**Number of Tests**	**Share Stat. Sig. 5%**	**Significant at 5%**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Econ*Post07	-117.241	0.097	-0.164	121.797	0.095	-0.014
	(76.477)	(0.077)	(0.135)	(59.930)	(0.079)	(0.106)
Economics	92.032	-0.079	0.173	-63.840	-0.062	-0.063
	(29.266)	(0.077)	(0.128)	(55.797)	(0.062)	(0.074)
Post 2007	-50.607	-0.081	0.285	-114.004	-0.642	-0.053
	(222.181)	(0.086)	(0.130)	(63.534)	(0.128)	(0.140)
Observations	322	54,565	3,339	319	22,716	1,412
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Reported Sig. by Means of Eye-Catchers	No	Yes	Yes	No	Yes	No
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes
Caliper	-	-	[1.81, 2.11]	-	-	[1.81, 2.11]
R^2	0.130	0.810	0.608	0.155	0.754	0.589
Adj. R^2	0.036	0.810	0.601	0.041	0.753	0.576

Models 1-3 report regression results from dissertations only. Models 4-6 report regression results from follow-up papers only. In all Models, we apply OLS regression with the outcome variable being the count of test statistics per dissertation or follow-up paper in Models 1 and 4. The outcome variable in Models 2-3 and 4-5 is an indicator variable for at least 5 percent statistical significance. Models 3 and 6 consider only test statistics inside a 0.150 caliper around the 1.96 *z*-value, i.e., absolute *z*-values between 1.81 and 2.11. Imprecise *z*-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with Post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Year FE and Region FE were kept fixed. Standard errors are clustered at the university-level in Models 1-3 and at author-level in Models 4-6.

Table D4—Main Regression like in Table 3.1 but considering only main tests and considering dissertation/follow-up paper weights

	Dissertations			Diss/Papers		
	Number of Tests (1)	**Share Stat. Sig. 5% (2)**	**Significant at 5% (3)**	**Number of Tests (4)**	**Share Stat. Sig. 5% (5)**	**Significant at 5% (6)**
Variables of Interest						
Graduate School	-18.841	0.011	0.006		-0.010	
	(24.376)	(0.020)	(0.052)		(0.020)	
Mandatory Supervision Agenda	71.702	-0.003	-0.047		-0.004	
	(45.067)	(0.022)	(0.045)		(0.025)	
Paper				-82.123	0.004	0.081
				(10.528)	(0.023)	(0.076)
Observations	289	26,606	1,882	541	34,374	2,522
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Reported Signifiance as Control	No	Yes	No	No	Yes	No
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes
Caliper	-	-	[1.81, 2.11]	-	-	[1.81, 2.11]
R^2	0.158	-	-	0.163	-	-
Adj. R^2	0.071	-	-	0.117	-	-
R^2 tjur	-	0.773	0.020	-	0.741	0.025
RMSE	-	0.228	0.507	-	0.242	0.507
F	-	87.746	3.893	-	148.038	5.832

Models 1-3 report regression results from dissertations only. Models 4-6 report regression results from dissertations and follow-up papers. In Models 1 and 4, we apply OLS regression, with the outcome variable being the count of test statistics per dissertation. Models 2-3 and 5-6 report average marginal effects from logit regressions, with the outcome variable in Models 2 and 5 being an indicator variable for at least 10 percent significance and an indicator variable for at least 5 percent significance in Models 3 and 6. Models 3 and 6 consider only test statistics inside a 0.150 caliper around the 1.96 *z*-value, i.e., absolute *z*-values between 1.81 and 2.11. Imprecise *z*-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with Post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Year FE and Region FE were kept fixed. We weigh observations by including the inverse of the number of tests in the dissertation/follow-up study. Standard errors are clustered at the university level in Models 1-3 and at the author level in Models 4-6.

Table D5—Main Regression like in Table 3.1 but without control variables.

	Dissertations			Diss/Papers		
	Number of Tests (1)	**Share Stat. Sig. 5% (2)**	**Significant at 5% (3)**	**Number of Tests (4)**	**Share Stat. Sig. 5% (5)**	**Significant at 5% (6)**
Variables of Interest						
Graduate School	-1.943	-0.033	0.007	1.195	-0.044	0.018
	(21.736)	(0.032)	(0.021)	(13.922)	(0.027)	(0.022)
Mandatory Supervision Agenda	45.558	-0.075	0.011	20.376	-0.079	-0.016
	(32.540)	(0.034)	(0.020)	(15.922)	(0.036)	(0.028)
Paper				-80.378	0.022	0.025
				(9.957)	(0.022)	(0.024)
Observations	289	34,853	2,275	543	45,412	2,952
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Caliper	-	-	[1.81, 2.11]	-	-	[1.81, 2.11]
R^2	0.008	-	-	0.086	-	-
Adj. R^2	0.001	-	-	0.081	-	-
R^2 tjur	-	0.004	0.000	-	0.006	0.001
RMSE	-	0.498	0.500	-	0.498	0,500
F	-	72.479	0.120	-	86.615	0.855

Models 1-3 report regression results from dissertations only. Models 4-6 report regression results from dissertations and follow-up papers. In Models 1 and 4, we apply OLS regression, with the outcome variable being the count of test statistics per dissertation. Models 2-3 and 5-6 report average marginal effects from logit regressions, with the outcome variable in Models 2 and 5 being an indicator variable for at least 5 percent significance and an indicator variable for at least 5 percent significance in Models 3 and 6. Models 3 and 6 consider only test statistics inside a 0.150 caliper around the 1.96 *z*-value, i.e., absolute *z*-values between 1.81 and 2.11. Imprecise *z*-values were removed following the approach of Kranz and Pütz (2022). Year FE and Region FE were kept fixed. Standard errors are clustered at the university level in Models 1-3 and at the author level in Models 4-6.

Table D6—Logit regression considering only dissertations that never produced a paper

	Dissertations		
	Number of Tests	**Share Stat. Sig. 5%**	**Significant at 5%**
	(1)	**(2)**	**(3)**
Graduate School	-17.324	0.005	0.076
	(33.129)	(0.015)	(0.034)
Mandatory Supervision Agenda	108.487	-0.059	0.032
	(54.738)	(0.018)	(0.063)
Observations	181	22,728	1,437
Year FE	Yes	Yes	Yes
Region FE	Yes	Yes	Yes
Reported Sig. as Control	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes
Caliper	-	-	[1.81, 2.11]
R^2	0.202	-	-
Adj. R^2	0.073	-	-
R^2 tjur	-	0.778	0.522
RMSE	-	0.235	0.345
F	-	68.362	10.276

Models 1-3 report regression results from dissertations only. In Model 1, we apply OLS regression, with the outcome variable being the count of test statistics per dissertation. Models 2-3 report average marginal effects from logit regressions, with the outcome variable in Model 2-3 being an indicator variable with at least 5 percent significance. Model 3 considers only test statistics inside a 0.150 caliper around the 1.96 *z*-value, i.e., absolute *z*-values between 1.81 and 2.11. Imprecise *z*-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with Post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Year FE and Region FE were kept fixed. Standard errors are clustered at the university level.

Table D7—Cumulative dissertation chapters that got published and their follow-up papers.

	Dissertations			Diss/Papers		
	Number of Tests	Share Stat. Sig. 5%	Significant at 5%	Number of Tests	Share Stat. Sig. 5%	Significant at 5%
	(1)	(2)	(3)	(4)	(5)	(6)
Variables of Interest						
Graduate School	-28.333	0.015	-0.003	-7.212	-0.006	0.079
	(56.674)	(0.024)	(0.063)	(40.177)	(0.017)	(0.039)
Mandatory Supervision Agenda	-63.084	-0.006	-0.282	-6.362	-0.047	-0.104
	(113.101)	(0.028)	(0.060)	(47.260)	(0.022)	(0.050)
Paper				-65.054	0.005	0.004
				(20.083)	(0.018)	(0.017)
Observations	108	13,036	848	269	24,043	1.494
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes
Caliper	-	-	[1.81, 2.11]	-	-	[1.81, 2.11]
R^2	0.105	-	-	0.088	-	-
Adj. R^2	-0.088	-	-	0.014	-	-
R^2 tjur	-	0.868	0,046	-	0.830	0.657
RMSE	-	0.179	0.488	-	0.205	0.294
F	-	16.405	0.965	-	68.486	14.247

In the dissertation-only data for Models 1-3, we keep only cumulative dissertation chapters that were later published as an empirical follow-up paper. In the overall data for Models 4-6, we added follow-up papers to the analysis. In Models 1 and 4, we apply OLS regression, with the outcome variable being the count of test statistics per cumulative dissertation chapter or follow-up paper. Models 2-3 and 5-6 report average marginal effects from logit regressions, with the outcome variable in Models 2 and 5 being an indicator variable for at least 10 percent statistical significance and an indicator variable for at least 5 percent statistical significance in Models 3 and 6. Models 3 and 6 consider only test statistics inside a 0.150 caliper around the 1.96 *z*-value, i.e., absolute *z*-values between 1.81 and 2.11. Imprecise *z*-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with Post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Year FE and Region FE are always included. Standard errors are clustered at the university-level in Models 1-3 and at author-level in Models 4-6.

Table D8—Cumulative dissertation chapters that were published and their follow-up papers

	Follow-Up Papers		
	Number of Tests	**Share Stat. Sig. 5%**	**Significant at 5%**
	(1)	**(2)**	**(3)**
Variables of Interest			
Graduate School	-59.928	-0.043	-0.002
	(38.219)	(0.023)	(0.039)
Mandatory Supervision Agenda	-5.485	-0.016	0.041
	(37.756)	(0.028)	(0.057)
After Defense	17.953	-0.008	-0.005
	(23.994)	(0.022)	(0.058)
Observations	161	13,181	947
Year FE	Yes	Yes	Yes
Region FE	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes
Caliper	-	-	[1.81, 2.11]
R^2	0.245	-	-
Adj. R^2	0.105	-	-
R^2 tjur	-	0.800	0.607
RMSE	-	0.222	0.313
F	-	36.072	8.710

We consider only follow-up papers that we could match to a cumulative dissertation chapter. In Model 1, we apply OLS regression, with the outcome variable being the count of test statistics per cumulative dissertation chapter or follow-up study. Models 2-3 report average marginal effects from logit regressions, with the outcome variable in Model 2 being an indicator variable for at least 5 percent statistical significance and an indicator variable for at least 5 percent statistical significance in Model 3. Model 3 considers only test statistics inside a 0.150 caliper around the 1.96 z-value, i.e., absolute z-values between 1.81 and 2.11. Imprecise z-values were removed following the approach of Kranz and Pütz (2022). Control variables were selected with Post-double lasso from the list of control variables we pre-defined in the pre-analysis plan. Year FE and Region FE are always included. Standard errors are clustered at the university level in Model 1 and at the author level in Models 2-3.