

# Researcher Ranking and Reporting Bias: Evidence from Economics

Stephan B. Bruns (U Hasselt and U Kassel)  
Kilian Buehling (Weizenbaum Institute and FU Berlin)  
Guido Buenstorf (U Kassel)  
Valon Kadriu (U Kassel)  
Andreas Rehs (Deutsche Bundesbank)



Leibniz Open Science Day 2024  
Meta Perspectives in Social Sciences  
November 25, 2024

# 1. Motivation

# Selective reporting contributes to the replication crisis in empirical research

- Selective reporting = likelihood that empirical results are published depends on their statistical significance → “reporting bias”



## Unreliable research Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

“I SEE a train wreck looming,” warned Daniel Kahneman, an eminent psychologist, in an open letter last year. The premonition concerned research on a phenomenon known as “priming”. Priming



Open access, freely available online

### Essay

## Why Most Published Research Findings Are False

John P. A. Ioannidis

### Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection

factors that influence this problem and some corollaries thereof.

### Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a  $p$ -value less than 0.05. Research is not most appropriately represented

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is  $R/(R+1)$ . The probability of a study finding a true relationship



PLoS Medicine | www.plosmedicine.org

0696

August 2005 | Volume 2 | Issue 8 | e124

interest and prejudice; and when more

truly exists reflects the type 1 error

# Selective reporting and competition

- Reporting bias may derive from various practices
  - Reviewers/editors favor statistically significant results → publication bias
  - Authors less likely to submit statistically insignificant results → file-drawer problem
  - Authors engage in practices that lead to inflated statistical significance → p-hacking
- Competition may increase prevalence of reporting bias
  - Link from competitive pressure to questionable research practices has often been suggested (Necker, 2014; Martin, 2016)
  - But scant empirical evidence to support this conjecture (Fanelli, 2020)
  - Specifically, lack of evidence regarding competition and selective reporting

# Ranking as a form of competition

- Competition is cornerstone of research system
  - However, various forms and arenas of competition
- “Quantified” competition based on indicators
  - Common metrics make performance comparable (Espeland & Stevens, 2008)
- Ranking: institutionalized comparison of actors
  - Ranking as a form of quantified competition that facilitates comparison and makes it visible (Espeland & Sauder, 2007)
  - Ranking may induce behavioral changes and “produce competition” (Brankovic et al., 2018)
    - Reactivity = “the idea that people change their behavior in reaction to being evaluated, observed, or measured” (Espeland & Sauder, 2007, p. 1; see already Campbell, 1957)



# Our study

- First systematic quantitative study to explore reactivity of an individual-level publication-based ranking
  - Focusing on selective reporting as a specific form of reactivity
  - Using difference-in-differences design in quasi-experimental setting (Blanco-Perez & Brodeur, 2020; Askarov et al., 2023)
- Note: our design explores effects of individual-level ranking, NOT effects of (quantified) competition more generally

## 2. Empirical context

# Handelsblatt ranking of economists (1)

- Individual-level (plus department level)

- Established in 2006-2007;  
since then repeated regularly
- Coverage: German-speaking countries
- Built on *Forschungsmonitoring.org*; after 2021 *Wirtschaftswoche* as media outlet

- Publication-based

- Indicator based on journal reputation  
(details have varied over time)
- No consideration of individual citation counts  
or activities other than publishing

HANDELSBLATT-RANKING VOLKSWIRTSCHAFTSLEHRE 2011

Top-100 Aktuelle Forschungsleistung (seit 2007)

Rang 2011	Rang 2010	Name	Universität	Alter	Fach	Punkte VWL 2011	Punkte A+	Punkte A und A+	Punkte / Publikation
1	1	Roman Inderst	Frankfurt / Main Uni	41	Industrieökonomie, Bankbetriebslehre & Finanzierung	9.98	2.5	8	0.31
2	4	Peter Egger	Zürich ETH	41	Internationale Ökonomie	7.02		3.15	0.11
3	2	Ernst Fehr	Zürich Uni	55	Experimentelle Wirtschaftsforschung	6.82	2.99	6.14	0.24
4	5	Marcel Fratzscher	Frankfurt EZB	40	Internationale Ökonomie, angewandte Makroökonomie	6.21	0.5	3.75	0.17
5	3	Matthias Sutter	Innsbruck Uni	42	Experimentelle Wirtschaftsforschung	5.67	1.5	4	0.19



# Handelsblatt ranking of economists (2)

- Introduction was exogenous event for most economists
- Very little controversy in economics (other than about methods)
  - Endorsement by *Verein für Socialpolitik* in 2007 (Hofmeister & Ursprung, 2008)
  - Widespread use in hirings and funding decisions (Berlemann & Haucap, 2015)

- Highly visible in German economics community (and beyond)

- Effect on reporting bias?

## Siegener Volkswirt unter den Top 100

Professor Dr. Thushyanthan Baskaran von der Universität Siegen zählt zu den forschungsstärksten deutschsprachigen Volkswirten. Das zeigt das aktuelle VWL Ranking der Zeitung „Handelsblatt“.

Göttingen / Handelsblatt-Ranking

10:56 Uhr / 04.10.2019

**Handelsblatt-Ranking: Gute Werte für  
Göttinger Ökonomen**

### 3. Analysis (still ongoing)

# Sample

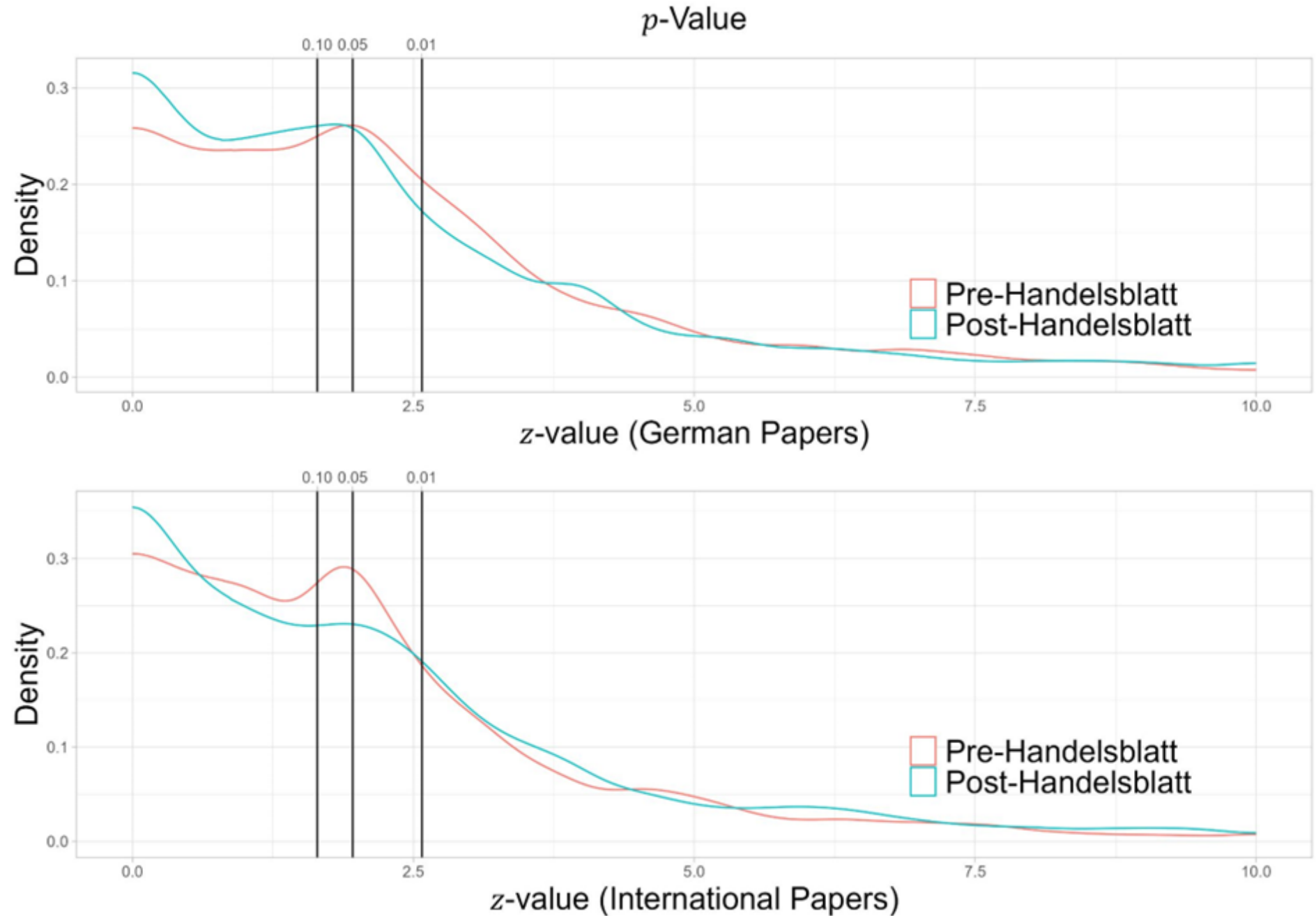
- Matched paper pairs (Germany-based authors vs. others)
  - Treatment articles: 1+ co-author in Germany; published in EconLit journals (2001-2014)
    - Treatment assumed to start in 2008 (reflects publication lags)
  - International control articles from same journal and issue; before/after focal German article (Fanelli et al., 2015)
    - Matching of paper pairs complicated by heterogeneity (theoretical papers; different methods)
  - Final sample: 190 closed paper pairs reporting results from hypothesis tests (90 pre-Handelsblatt; 100 post-Handelsblatt)
- Use information about (absolute) z-values to trace selective reporting
  - Manual extraction of test statistics, coefficients, standard errors etc. and other meta data by trained RAs (with supervision and resolution of ambiguous cases by co-authors)
  - In total, 19,191 z-values are calculated from reported statistical information
    - Substantial variation in number of z-values per paper (min.: 2; max.: 811)

# Share all of z-values rejecting null hypothesis

- Has tendency to report statistically significant results increased?
  - All z-values from matched paper pairs considered
  - Overall share of statistically significant results actually decreased in Germany (but starting from higher initial levels)

Significance Level	Region	z-value (2001-2007)	z-value (2008-2014)	Share Above (2001-2007)	Share Above (2008-2014)
10%	German	3681	6859	0.641	0.592
10%	International	3809	4842	0.596	0.591
5%	German	3681	6859	0.566	0.523
5%	International	3809	4842	0.500	0.528
1%	German	3681	6859	0.424	0.390
1%	International	3809	4842	0.364	0.400

# Density of (absolute) z-values



# Binomial tests for various caliper sizes

- Caliper test (Gerber & Malhotra, 2008)
  - Share of significant z-values in various ranges above/below conventional significance thresholds (focus on 5% level, also analyzed 10% and 1%)
- Are imbalances statistically significant?
  - Confidence intervals based on bootstrapping (resampling on paper level) to account for non-independence of observations
- Did they get more pronounced after 2007?
  - Germany vs. international papers?

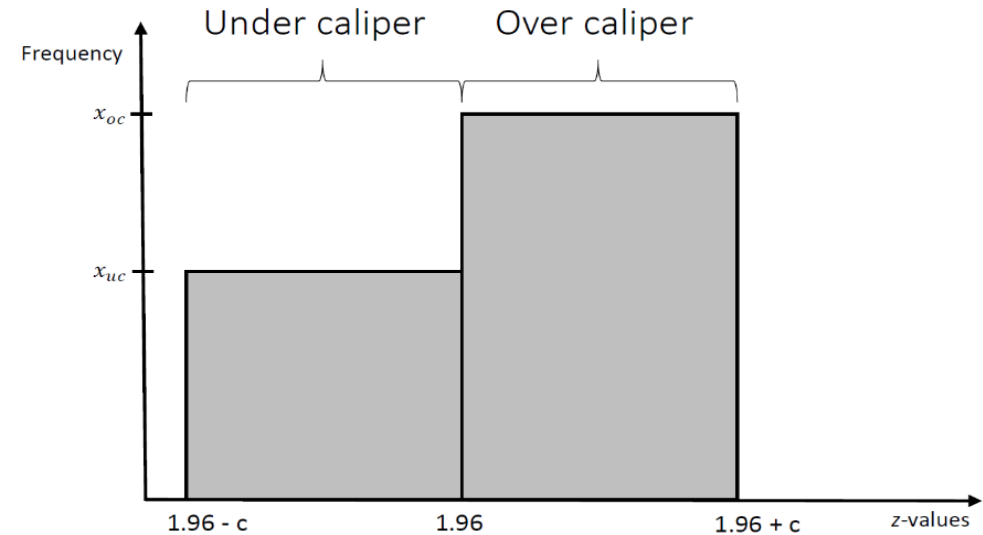


Figure: Bruns et al., 2019

# Binomial tests (5% level)

	All		German		International	
	Pre	Post	Pre	Post	Pre	Post
Caliper Size	0.150					
No. of Tests in Caliper	636	912	295	584	341	328
Under Caliper	281	380	120	224	161	156
Over Caliper	355	532	175	360	180	172
Binomial Probability	<b>0.558</b>	<b>0.583</b>	<b>0.593</b>	<b>0.616</b>	0.528	0.524
95% Conf. Interval	[0.514, 0.623]	[0.517, 0.624]	[0.541, 0.651]	[0.534, 0.667]	[0.466, 0.645]	[0.481, 0.567]
Caliper Size	0.100					
No. of Tests in Caliper	450	642	201	417	249	224
Under Caliper	203	239	86	142	117	97
Over Caliper	247	402	115	275	132	127
Binomial Probability	<b>0.549</b>	<b>0.627</b>	<b>0.572</b>	<b>0.659</b>	0.530	<b>0.567</b>
95% Conf. Interval	[0.502, 0.611]	[0.554, 0.670]	[0.504, 0.639]	[0.550, 0.710]	[0.466, 0.637]	[0.514, 0.619]
Caliper Size	0.050					
No. of Tests in Caliper	246	347	109	219	137	128
Under Caliper	108	99	46	52	62	47
Over Caliper	138	248	63	167	75	81
Binomial Probability	0.561	<b>0.715</b>	0.578	<b>0.763</b>	0.547	<b>0.633</b>
95% Conf. Interval	[0.494, 0.649]	[0.631, 0.768]	[0.483, 0.662]	[0.647, 0.813]	[0.453, 0.682]	[0.560, 0.699]

# Logistic regressions: approach

- Difference-in-differences framework (Blanco-Perez & Brodeur, 2020)

$$P(Y = 1|X) = F(\alpha + \beta_1 \textit{German} + \beta_2 \textit{Post2007} + \beta_3 \textit{German} * \textit{Post2007} + \beta_4 X)$$

- $Y = 1$  for z-values in over-caliper implying statistical significance
- Interaction term  $\textit{German} * \textit{Post2007}$  captures *Handelsblatt* effect
  - Standard errors clustered at article level
  - With / without controls
  - No weighting / weighting at paper level / weighting at paper-pair level
  - Linear probability models as robustness checks



# Logistic regressions: Baseline model (5% level)

- No controls, unweighted

	Significant at the 5% level					
	All observations (1)	0.500 Caliper (2)	0.300 Caliper (3)	0.150 Caliper (4)	0.100 Caliper (5)	0.050 Caliper (6)
GermanPost07	-0.063 (0.069)	-0.052 (0.049)	-0.015 (0.076)	0.026 (0.086)	0.053 (0.091)	0.109 (0.109)
german	0.063 (0.057)	0.032 (0.043)	0.048 (0.064)	0.065 (0.064)	0.041 (0.064)	0.027 (0.085)
post07	0.023 (0.056)	0.029 (0.043)	0.028 (0.064)	-0.003 (0.059)	0.036 (0.059)	0.079 (0.083)
r2.tjur	0.002	0.001	0.002	0.007	0.012	0.036
rmse	0.499	0.500	0.499	0.493	0.488	0.468
nobs	19190	4395	2846	1549	1091	593
F	10.293	0.942	2.067	3.578	4.194	6.850

Notes: This table reports average marginal effects from a logistic regression. The outcome variable is a dummy taking the value 1 if the corresponding calculated z-value is statistically significant at the 5% level. In Models 2-6 we reduce our sample to values around a 0.500, 0.300, 0.150, 0.100, and 0.050 caliper respectively around the 1.96 z-value. Clustered standard errors at the article level are reported in brackets. \*  $p < 0.10$ ; \*\*  $p < 0.05$ ; \*\*\*  $p < 0.01$ .

# Logistic regressions: full model (5% level)

- With controls, weighted at paper-pair level
  - Controls: # tests, female co-author, # authors, # countries, HB journal weight, year FEs
  - Note: Post07 dummy absorbed in year FEs

	Significant at the 5% level					
	All observations (1)	0.500 Caliper (2)	0.300 Caliper (3)	0.150 Caliper (4)	0.100 Caliper (5)	0.050 Caliper (6)
GermanPost07	0.007 (0.052)	0.047 (0.051)	0.121* (0.064)	0.163** (0.081)	0.090 (0.092)	0.127 (0.120)
german	0.015 (0.037)	-0.040 (0.036)	-0.073 (0.047)	-0.097 (0.061)	-0.062 (0.068)	-0.041 (0.084)

# Logistic regressions: overview (5% level)

	German*Post07 (AME)					
	<b>All observations (1)</b>	<b>0.500 Caliper (2)</b>	<b>0.300 Caliper (3)</b>	<b>0.150 Caliper (4)</b>	<b>0.100 Caliper (5)</b>	<b>0.050 Caliper (6)</b>
No controls unweighted	-0.063 (0.069)	-0.052 (0.049)	-0.015 (0.076)	0.026 (0.086)	0.053 (0.091)	0.109 (0.109)
LPM coefficient No cont., unweighted	-0.063 (0.070)	-0.052 (0.049)	-0.016 (0.076)	0.026 (0.086)	0.050 (0.092)	0.099 (0.118)
No controls paper-level weights	-0.063 (0.069)	-0.052 (0.049)	-0.015 (0.076)	0.026 (0.086)	0.053 (0.091)	0.109 (0.109)
No controls paper-pair weights	-0.007 (0.055)	0.017 (0.054)	0.090 (0.065)	0.127 (0.084)	0.084 (0.099)	0.112 (0.124)
With controls paper-level weights	-0.010 (0.052)	-0.007 (0.040)	0.010 (0.052)	0.010 (0.067)	0.010 (0.072)	0.134 (0.086)
With controls paper-pair weights	0.007 (0.052)	0.047 (0.051)	0.121* (0.064)	0.163** (0.081)	0.090 (0.092)	0.127 (0.120)

Notes: This table reports average marginal effects from a logistic regression. The outcome variable is a dummy taking the value 1 if the corresponding calculated z-value is statistically significant at the 5% level. In Models 2-6 we reduce our sample to values around a 0.500, 0.300, 0.150, 0.100, and 0.050 caliper respectively around the 1.96 z-value. Clustered standard errors at the article level are reported in brackets. \*  $p < 0.10$ ; \*\*  $p < 0.05$ ; \*\*\*  $p < 0.01$ .

# Logistic regressions: additional analyses

- International matches with at least one author from US / UK
    - Rationale: no system-level change in US; RAE/REF in UK since 1986
  - Anticipation and lagged effects
    - Germany\*Post2006; Germany\*Post2008
  - Exclusion of transition period 2007-2009 (publication lags)
  - Absolute z-values as outcomes (Askarov et al., 2023)
- Interactions mostly positive but close to 0 and statist. insignificant

## 5. Discussion

# Discussion

- No robust evidence that *Handelsblatt* ranking led to more selective reporting in German economics
  - Consistent with meta-analysis by Fanelli et al. (2017)
- Again: we explore effects of ranking, not of competition as such
  - Despite broad interest in reactivity, surprisingly little prior evidence exists
- Possible interpretation: conditional on prior acceptance of (quantified) competition, ranking has limited effect on practices
  - (Quantified) competition in economics not limited to Germany
    - International trend toward quantitative research assessment
    - Published output comparisons of European econ departments (Combes & Linnemer, 2003)
    - Publication-based hiring decisions (Graber & Wälde, 2008)
  - *Handelsblatt* ranking was far more controversial in *Betriebswirtschaftslehre*

Thank you!

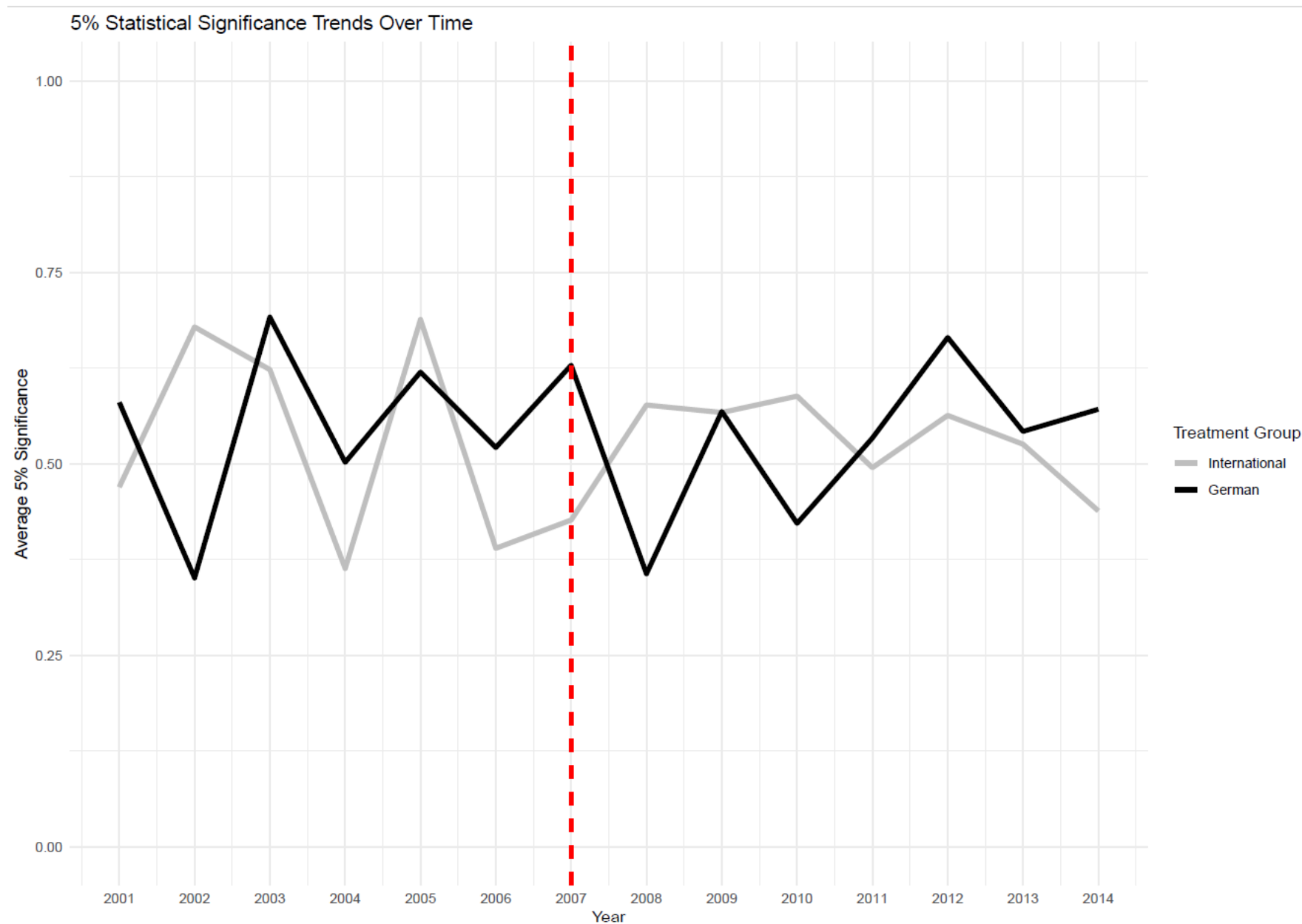
# Backup



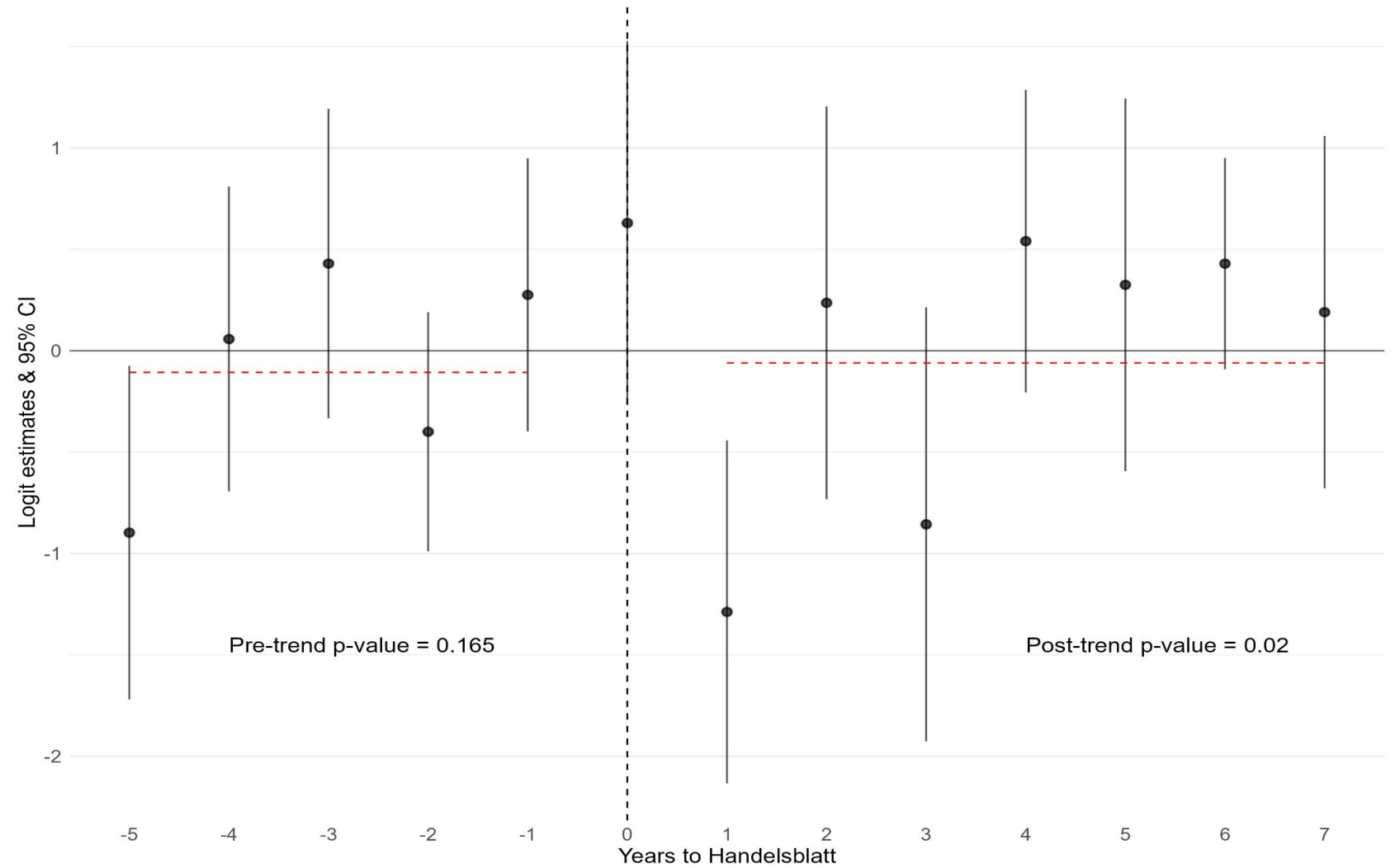
**Table 1.** *Descriptive Statistics*

	<b>No. of papers</b>	<b>No. of z-values</b>	<b>Mean</b>	<b>Min</b>	<b>Max</b>
Papers from closed Pairs	380	19,191	50.50	2	811
German Papers from closed Pairs	190	10,540	55.47	2	811
Pre 2008	180	7,490	41.61	2	437
Post 2007	200	11,701	58.51	2	811
Test Stat Reporting					
z-value	32	1,463	45.72	1	216
p-value	30	1,340	44.67	3	437
t-value	122	5,142	42.15	2	336
Coefficient and Standard Error	206	11,246	54.59	1	811
Author Count					
Author Count = 1	96	4,224	44.00	4	290
Author Count > 1	284	14,967	52.70	2	811
Number of different countries					
Number of different countries = 1	230	9,603	41.75	2	290
Number of different countries > 1	148	9,575	64.70	2	811
Male Quota					
Male Quota = 100%	241	11,653	48.35	2	811
Male Quota < 100%	138	7,535	54.60	3	437

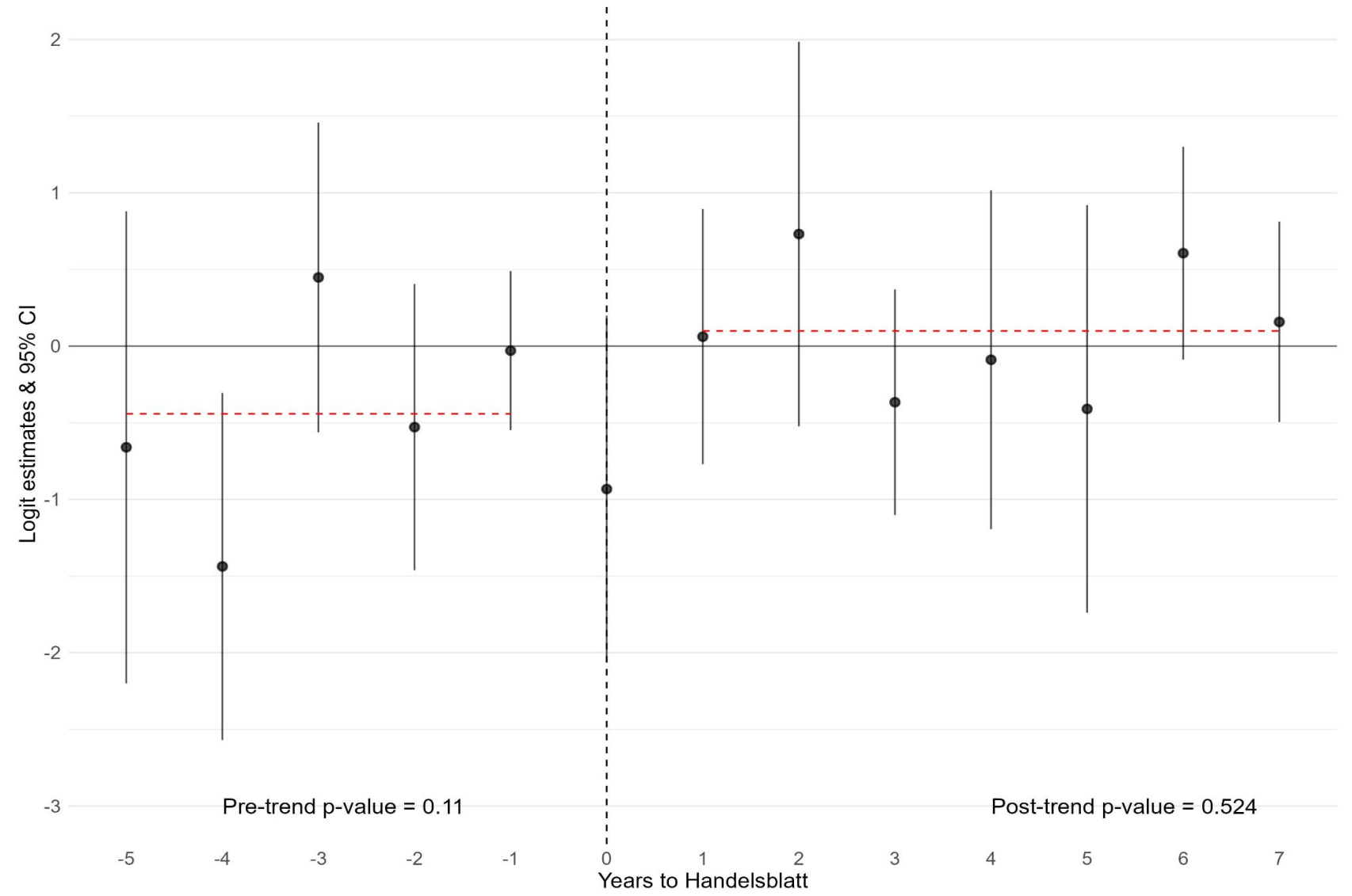
Common  
trends?  
( $p > 0.5$ )



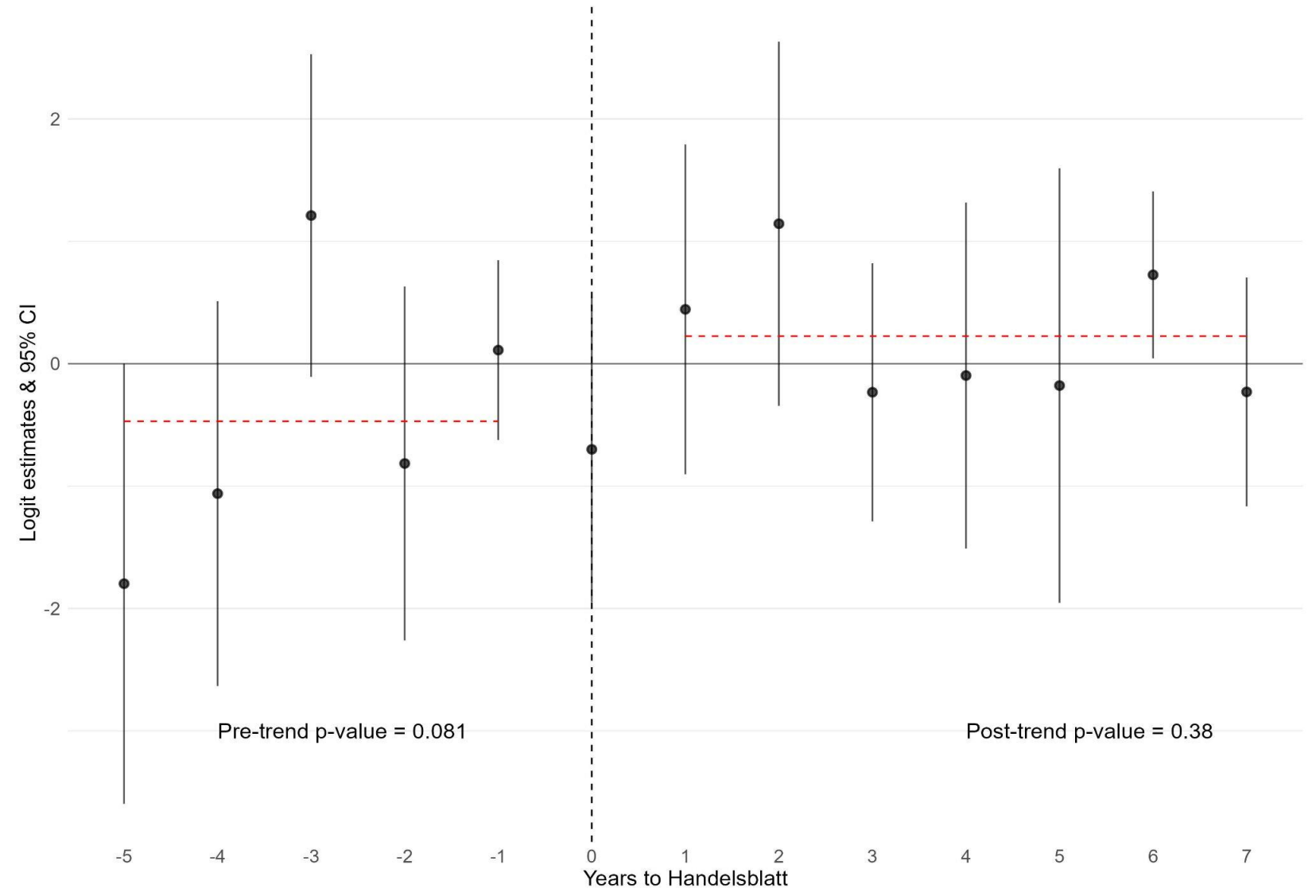
Common  
trends?  
(German\*T)



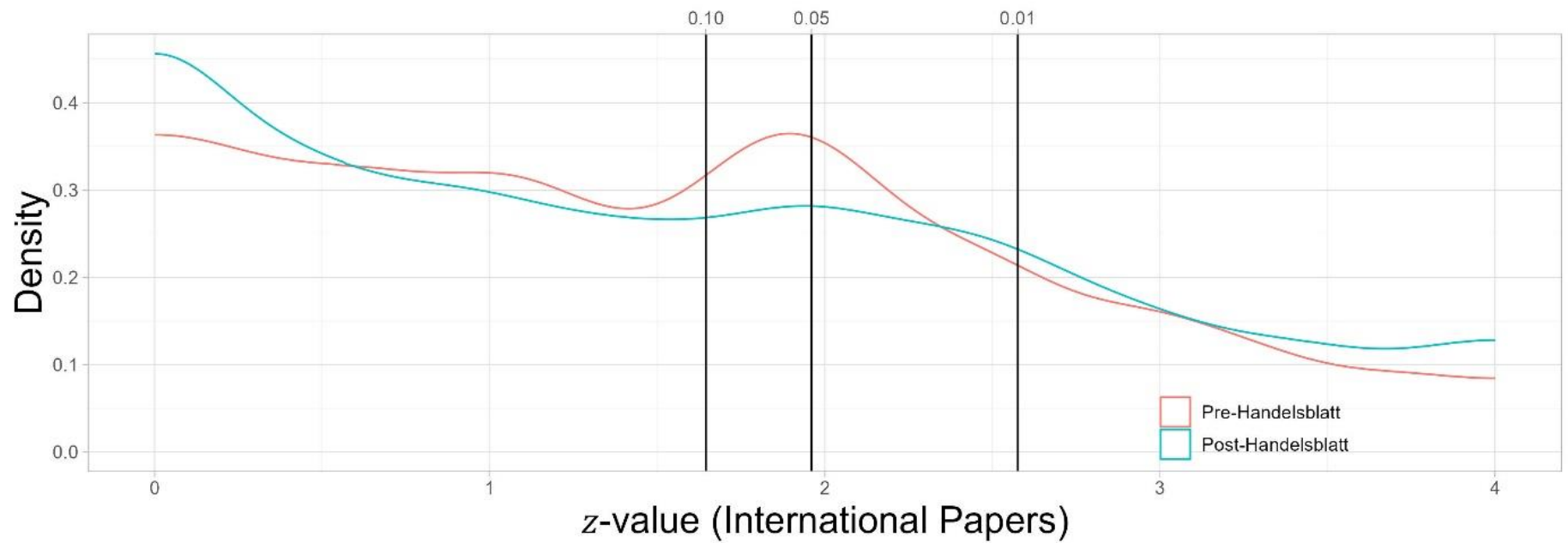
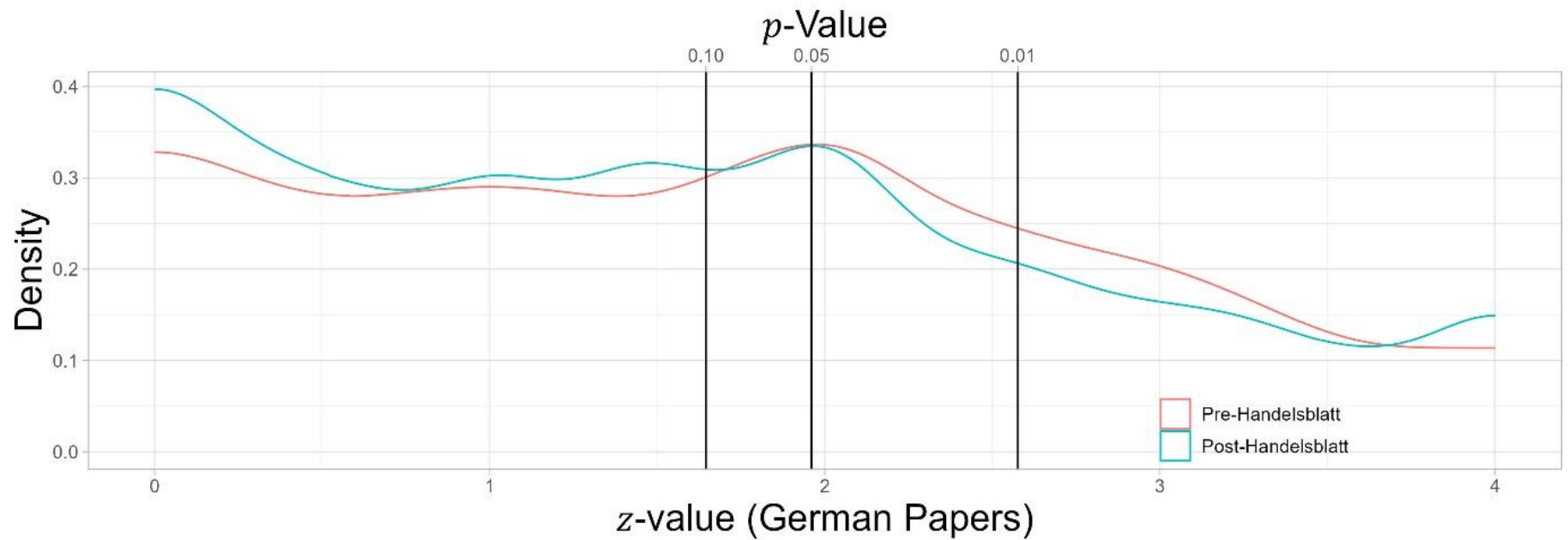
Common  
trends?  
(German\*T)  
(c = 0.300)



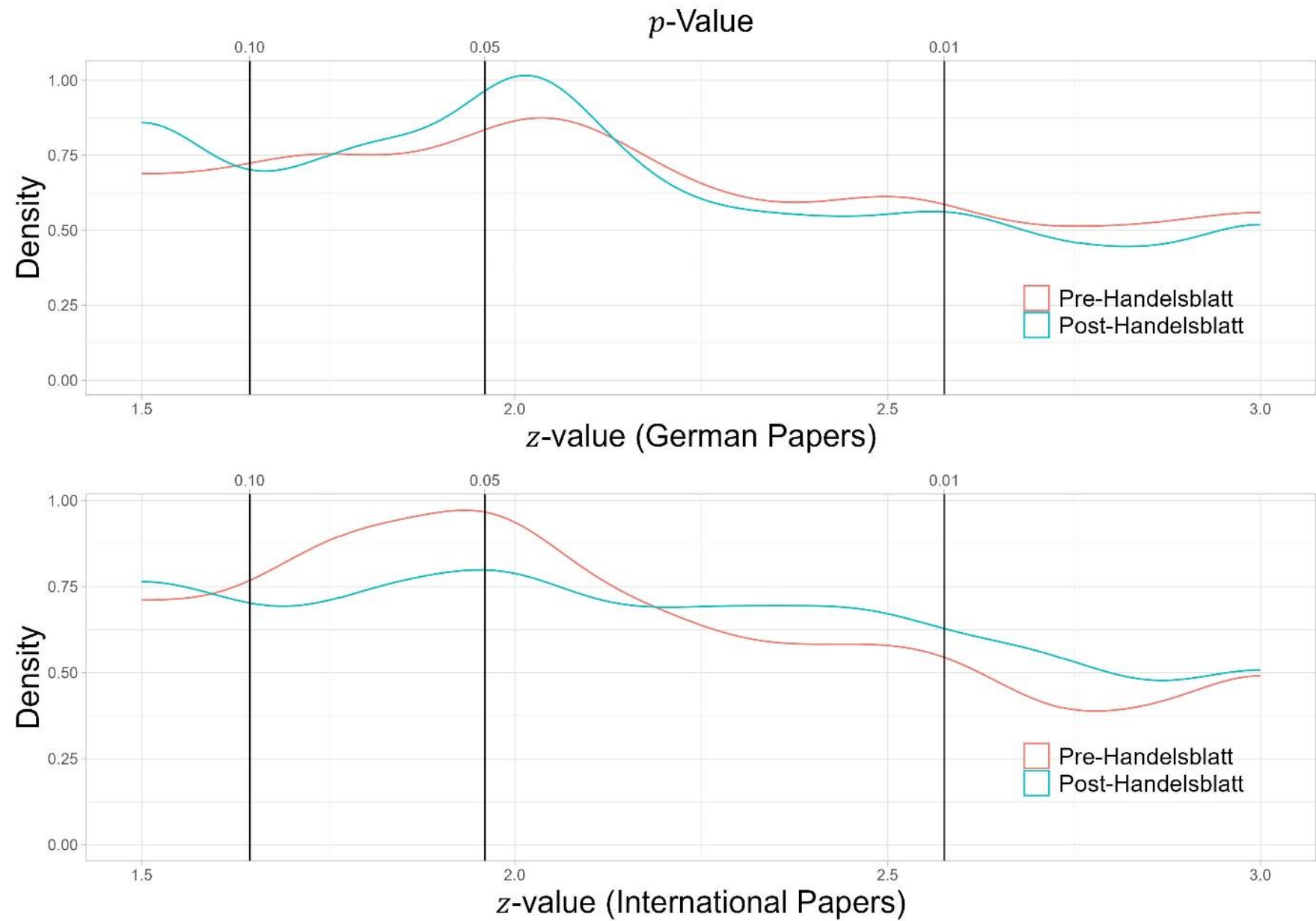
Common  
trends?  
(German\*T)  
(c = 0.150)



Zooming  
in...






# Density of (absolute) z-values



# Handelsblatt ranking of economists (3)

- From the very first edition (2005; based on a different methodology)

## Die Medienstars

1	1 214 Zitate	Alter
	<b>Hans-Werner Sinn (57)</b>	Der Ifo-Präsident ist Medienliebling Nr. 1 unter den deutschen Ökonomen - nicht erst seit seinem Sachbuch-Bestseller „Ist Deutschland noch zu retten?“
2	732 Zitate	
	<b>Bert Rürup (61)</b>	Wichtigster Regierungsberater und seit März auch Vorsitzender des Sachverständigenrats, vor allem bei Themen rund um die soziale Sicherung ein gefragter Interviewpartner.
3	570 Zitate	
	<b>Karl Lauterbach (42)</b>	Das vehemente Eintreten für die Bürgerversicherung machte den Kölner Gesundheitsökonom bundesweit bekannt - und zum zentralen Widersacher Rürups.

11	264 Zitate	
	<b>Horst Siebert (67)</b>	Der pensionierte Präsident des Kieler Instituts für Weltwirtschaft und ehemalige Wirtschaftsweise ist auch im Ruhestand ein viel zitierter Ökonom.
12	243 Zitate	
	<b>Joachim Scheide (55)</b>	Der Konjunkturchef des Kieler Instituts für Weltwirtschaft ist ein überzeugter Vertreter der angebotsorientierten Wirtschaftspolitik.
13	207 Zitate	
	<b>Gernot Nerb (62)</b>	Der Bereichsleiter Branchenforschung des Ifo-Instituts war bis Mitte 2004 für den Ifo-Index verantwortlich und kommentiert ihn auch heute regelmäßig.
14	184 Zitate	
	<b>Rolf Peffekoven (66)</b>	Der Mainzer Finanzwissenschaftler war von 1991 bis 2001 Mitglied des Sachverständigenrates und wird von den Medien vor allem zur Steuer- und Haushaltspolitik befragt.
15	151 Zitate	
	<b>Martin Hellwig (56)</b>	Der Direktor des Bonner Max-Planck-Instituts zur Erforschung