
RUB
PÄDAGOGISCHE
PSYCHOLOGIE

RUHR-UNIVERSITÄT BOCHUM

Institut für Erziehungswissenschaft

RUB PÄDAGOGISCHE PSYCHOLOGIE

RUB

Diverses mit “P”: Präregistrierung, Power, p-Hacking, Peer Review und Publikationsbias

“It was a quest for aesthetics, for beauty — instead of the truth”

Diederik Stapel

Bhattacharjee, Y. (2013, April 26). The mind of a con man. *New York Times*. Retrieved from http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html

Die Stapel-Affäre

- Wenigstens 50 Fälle wissenschaftlichen Fehlverhaltens (*Levelt Committee*)
  www.commisielevelt.nl
- Datensätze verändert, manipuliert, und erfunden um in hochrangigen Journals zu publizieren (inkl *Science*)
  - Vermülle Umgebungen verstärken rassistische Vorurteile (Stapel & Lindenberg, 2011)
  - Verminderung rassistischer Vorurteile bei Richtern gegenüber Angeklagten (Lammers & Stapel, 2011)
- Betrug ein Dämpfer für die Sozialpsychologie, und auch ein wenig für die Gesellschaft insgesamt

Questionable Research Practices (QRPs)

- Schwerwiegendes Fehlverhalten in allen wissenschaftlichen Disziplinen eher selten
  - Anstrengend Kartenhaus aufrecht zu erhalten
  - Die “Community” wird Betrug (zumindest schwere Fälle) vermutlich nicht verzeihen
- Grauzone: QRPs/*p*-hacking
  - Höhere Akzeptanz unter Wissenschaftlern

Was sind QRPs?

Selektives Berichten

- Nicht alle gemessenen Variablen berichten (***outcome switching***)
- Nicht alle ausprobierten Analysestrategien berichten
- Nicht alle experimentellen Bedingungen berichten
- Nicht alle Studien berichten

Methodische Flexibilität

- Zusätzliche Datenerhebung nach dem Signifikanztest (***peeking***)
- Datenerhebung stoppen nach Signifikanztest (***optional stopping***)
- Ausreißerbereinigung nach Signifikanztest
- p-Werte abrunden
- Erst analysieren, dann Hypothesen aufschreiben (***HARKing***)
- Einseitig testen wenn $0.05 < p < 0.10$

Was sind QRPs?

Selektives Berichten

- Nicht alle gemessenen Variablen berichten (***outcome switching***)
- Nicht alle ausprobierten Analysestrategien berichten
- Nicht alle experimentellen Bedingungen berichten
- Nicht alle Studien berichten

Methodische Flexibilität

- Zusätzliche Datenerhebung nach dem Signifikanztest (***peeking***)
- Datenerhebung stoppen nach Signifikanztest (***optional stopping***)
- Ausreißerbereinigung nach Signifikanztest
- p-Werte abrunden
- Erst analysieren, dann Hypothesen aufschreiben (***HARKing***)
- Einseitig testen wenn $0.05 < p < 0.10$

Multiple Analysestrategien

- AVs = oft Aggregat aus mehreren Messungen
- Strategien zur Aggregierung potenziell unendlich
  - Kombinationen von self-report Items einer Skala
  - Verrechnung verschiedener Trials zu einem Index
  - Regions of Interest (ROI)
- Selbstbetrug ziemlich leicht (man findet immer opportune *Best Practice*-Empfehlung)

Multiple Analysestrategien: Beispiel

- **Competitive Reaction Time Task (CRTT)**
- Populärstes Verfahren zur Messung aggressiven Verhaltens im Labor
- Reaktionszeitspiel gegen anderen Probanden (mehrere Runden)
- Vor jeder Runde wird die Intensität eines ‘Noise Blasts’ festgelegt (= Aggression)
- Visuelles Signal -> so schnell wie möglich Leertaste drücken
- Verlierer hört Noise Blast mit den Einstellungen des Gegners

Multiple Analysestrategien: Beispiel

- Mean volume (Anderson & Carnagey, 2009)
- Mean volume after wins (Anderson & Dill, 2000)
- Mean volume after losses (Anderson & Dill, 2000)
- Mean volume x duration (Bartholow, Sestir, & Davis, 2005)
- Mean volume x $\sqrt{duration}$ (Carnagey & Anderson, 2005)
- Mean volume x $\log_e$(duration) (Lindsay & Anderson, 2000)
- Separate means for trials 2-9, 10-17, and 18-25 (Anderson et al., 2004)
- Sum of z(volume) and z(duration) (Sestir & Bartholow, 2010)
- Total high volume settings (Anderson & Carnagey, 2009)
- First trial volume (Bushman & Baumeister, 1998)
- …

Multiple Analysestrategien: Beispiel

Volume, # of high settings (3-4) in trials 1-24
Volume, # of high settings (3-4) in trials 25-48
Volume, # of high settings (6-8) in all trials (14)
Volume, # of high settings (6-8) in trials 1-42
Volume, # of high settings (6-8) in trials 43-84
Volume, # of high settings (7-10) in trials 1-15
Volume, # of high settings (7-10) in trials 16-25
Volume, # of high settings (8-10) in all trials (25)
Volume, # of high settings (8-10) in all trials (25), square-rooted
Volume, # of high settings (9-10) in all trials (25)
Volume, # of high settings (9-10) in all trials (33)
Volume, # of low settings (1-3) in trials 1-42
Volume, # of low settings (1-3) in trials 43-84
Volume, # of maximum settings (10) in all trials (25)
Volume, # of maximum settings (10) in all trials (30)
Volume, # of maximum settings (10) in all trials (5)
Volume, # of maximum settings (8) in trials 1-160
Volume, # of maximum settings (8) in trials 161-320
Volume, # of maximum settings (9) in all trials (12)
Volume, # of medium settings (4-5) in trials 1-42
Volume, # of medium settings (4-5) in trials 43-84
Volume, # of minimum settings (0) in all trials (12)
Volume, # of minimum settings (0) in all trials (30)
Volume, # of minimum settings (1) in all trials (25)
Volume, # of minimum settings (1) in trials 1-160
Volume, # of minimum settings (1) in trials 161-320
Volume, % of winning trials with maximum setting (8)
Volume, after losing, average of 12 trials
Volume, after losing, average of a variable number of trials
Volume, after winning following a prior loss, average of a variable number of trials
Volume, after winning, # of high settings (7-9) in 24 trials
Volume, after winning, average of 13 trials
Volume, after winning, average of 18 trials
Volume, after winning, average of 24 trials
Volume, after winning, average of a variable number of trials
Volume, after winning, average of trials 1-12
Volume, after winning, average of trials 13-24
Volume, after winning, average of trials 25-36
Volume, average of all trials (12)
Volume, average of all trials (14)
Volume, average of all trials (20)
Volume, average of all trials (24)
Volume, average of all trials (25)2
Volume, average of all trials (25), standardized
Volume, average of all trials (30)
Volume, average of all trials (5)
Volume, average of all trials (50)
Volume, average of trials 1-10
Volume, average of trials 1-160
Volume, average of trials 1-20
Volume, average of trials 1-24
Volume, average of trials 1-3
Volume, average of trials 1-42
Volume, average of trials 1-8
Volume, average of trials 10-17
Volume, average of trials 11-20
Volume, average of trials 14-19
Volume, average of trials 161-320
Volume, average of trials 17-24
Volume, average of trials 18-25
Volume, average of trials 2-19
Volume, average of trials 2-25
Volume, average of trials 2-41
Volume, average of trials 2-50
Volume, average of trials 2-7
Volume, average of trials 2-8 regressed on trial 1, residuals
Volume, average of trials 2-9
Volume, average of trials 20-25
Volume, average of trials 21-30
Volume, average of trials 21-40
Volume, average of trials 25-48
Volume, average of trials 3-9
Volume, average of trials 42-81
Volume, average of trials 43-84
Volume, average of trials 8-13
Volume, average of trials 9-16
Volume, first trial3
Volume, first trial in trials 1-160
Volume, first trial in trials 161-320
Volume, highest setting in all trials (30)
Volume, second trial
Volume, slope across all trials (25)
Duration, after losing, average of 12 trials, logarithmized
Duration, after winning, average of 13 trials, logarithmized
Duration, average of all trials (20)
Duration, average of all trials (25)
Duration, average of trials 1-8
Duration, average of trials 10-17
Duration, average of trials 17-24
Duration, average of trials 18-25
Duration, average of trials 2-9
Duration, average of trials 3-9
Duration, average of trials 9-16
Duration, first trial
Duration, second trial
Volume + Duration (mean), average of all trials (10)
Volume + Duration (mean), average of all trials (16)
Volume + Duration (mean), average of all trials (20)
Volume + Duration (mean), average of all trials (25)
Volume + Duration (mean), average of all trials (25), standardized
Volume + Duration (mean), average of all trials (30)
Volume + Duration (mean), average of trials 1-10
Volume + Duration (mean), average of trials 10-17
Volume + Duration (mean), average of trials 11-20
Volume + Duration (mean), average of trials 18-25
Volume + Duration (mean), average of trials 2-9
Volume + Duration (mean), average of trials 2-9, standardized
Volume + Duration (mean), average of trials 21-30
Volume + Duration (mean), first trial
Volume + Duration (mean), first trial, standardized
Volume + Duration (mean), second trial
Volume + Duration (sum), average of all trials (25), standardized
Volume + Duration (sum), average of all trials (30)
Volume + Duration (sum), average of all trials (8), standardized
Volume + Duration (sum), average of all trials (9), standardized
Volume + Duration (sum), average of trials 1-6
Volume + Duration (sum), average of trials 1-6, square-rooted
Volume + Duration (sum), average of trials 10-17
Volume + Duration (sum), average of trials 18-25
Volume + Duration (sum), average of trials 2-25, standardized
Volume + Duration (sum), average of trials 2-9
Volume + Duration (sum), average of trials 3-9, standardized
Volume + Duration (sum), average of trials 7-30
Volume + Duration (sum), first trial
Volume + Duration (sum), first trial, standardized
Volume + Duration (sum), first trial, standardized, increased by 10, logarithmized (base 10)
Volume + Duration (sum), second trial, standardized
Volume x Duration, # of high settings (80th percentile) in all trials (25)
Volume x Duration, # of high settings (85th percentile) in all trials (25)
Volume x Duration, # of high settings (90th percentile) in all trials (25)
Volume x Duration, average of all trials (25)
Volume x Duration, first trial, logarithmized
Volume x Duration, multiplied averages of all trials (25)
Volume x log(Duration), after losing, average of 13 trials
Volume x log(Duration), after winning, average of 12 trials
Volume x log(Duration), average of all trials (25)
Volume x log(Duration), linear contrasts across all trials (25)
Volume x log(Duration), quadratic contrasts across all trials (25)
Volume x √Duration, after losing, average of 4 trials
Volume x √Duration, average of all trials (25)
Any setting latency (# of trials until the first setting > 0)
Maximum setting latency (# of trials until maximum volume was set the first time)
Reaction time, average of all trials (25)
Setting time (seconds), average of all trials (25), logarithmized, winsorized
Setting time (seconds), average of trials 14-19
Setting time (seconds), average of trials 2-19
Setting time (seconds), average of trials 2-7
Setting time (seconds), average of trials 20-25
Setting time (seconds), average of trials 8-13
Setting time (seconds), first trial

Multiple Analysestrategien: Beispiel

Wie häufig sind QRPs?

Selektives Berichten

- Nicht alle gemessenen Variablen berichten (***outcome switching***)
- Nicht alle ausprobierten Analysestrategien berichten
- Nicht alle experimentellen Bedingungen berichten
- Nicht alle Studien berichten

Methodische Flexibilität

- Zusätzliche Datenerhebung nach dem Signifikanztest (***peeking***)
- Datenerhebung stoppen nach Signifikanztest (***optional stopping***)
- Ausreißerbereinigung nach Signifikanztest
- p-Werte abrunden
- Erst analysieren, dann Hypothesen aufschreiben (***HARKing***)
- Einseitig testen wenn $0.05 < p < 0.10$

Wie häufig sind QRPs?

Methodische Flexibilität – Selektives

Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling

Leslie K. John[1], George Loewenstein[2], and Drazen Prelec[3]
[1]Marketing Unit, Harvard Business School; [2]Department of Social & Decision Sciences, Carnegie Mellon University; and [3]Sloan School of Management and Departments of Economics and Brain & Cognitive Sciences, Massachusetts Institute of Technology

- p-Werte abrunden
- Erst analysieren, dann Hypothesen aufschreiben (***HARKing***)
- Einseitig testen wenn $0.05 < p < 0.10$

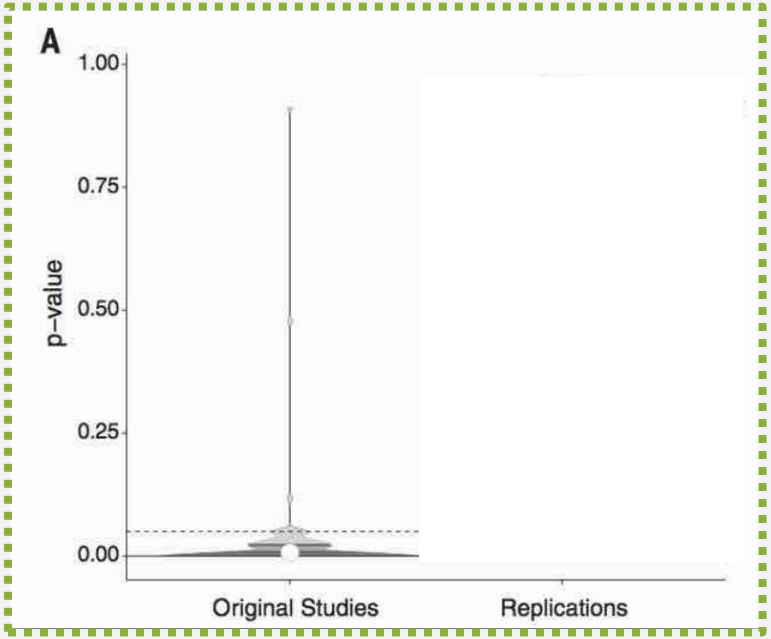
Welche Auswirkungen haben QRPs?

"QRPs are the steroids of scientific competition"

(John, Prelec, & Loewenstein, 2012)

- Erhöhen die Wahrscheinlichkeit, signifikante Ergebnisse für eine Hypothese zu finden
- Erhöhen künstlich die "Performanz" von Wissenschaftler im akademischen Wettbewerb um Geld und Aufmerksamkeit
- Integer arbeitende Wissenschaftler werden benachteiligt

Konsequenzen von p-Hacking



- Reproducibility Project: Psychology (OSC2015)
- 97 Replikationen
- 36% aller kritischer Tests statistisch signifikant (>98% bei den Originalstudien)

Einige relevante Dinge mit „P“

Transparenz erhöhen

- Peer Review
- Publizieren von Daten
- …

Freiheitsgrade einschränken

- **Preregistration**
- Poweranalysen
- *p*-Werte prüfen
- …

Preregistration

- **Formalisiert den idealen Ablauf wissenschaftlicher Forschung**
- Erlaubt Unterscheidung zwischen (dis-)konfirmatorischer & explorativer Forschung
- Erlaubt die Interpretation von *p*-Werten zum konfirmatorischen Hypothesentesten
- Schützt vor Selbstbetrug
- Schützt vor Reviewer #2
- Zwingt einen vorher über vieles nachzudenken

Preregistration: Formalisierung

Phase 1	Phase 2
Literaturüberblick Theorie Hypothesen Methoden Sampling Plan Analyseplan (+Skripte) Ausschlusskriterien Entscheidungsbaum **Paper schreiben bis zum Results Teil**	Datenerhebung Auswertung nach Plan aus Phase 1 Interpretation (nach Plan) ggf. explorative Auswertung **Rest des Papers schreiben**

Preregistration: Konfirmatorisch vs. Explorativ

Konfirmatorisch

- Testet Hypothesen
- Konkrete Vorhersagen
- Quantifizierung von Zusammenhängen
  - Erlaubt sinnvolle Stichprobenplanung!
- Festlegung auf statistische Operationalisierungen

Explorativ

- Generiert Hypothesen
- Vorrausetzung für zukünftige Vorhersagen
- Informationen für Quantifizierung von Zusammenhängen
- Ausprobieren möglicher statistischer Operationalisierungen

Einige relevante Dinge mit „P“

Transparenz erhöhen

Peer Review

Publizieren von Daten

…

Freiheitsgrade einschränken

Preregistration

Poweranalysen

p-Werte prüfen

…

Preregistration: Stichprobenplanung

- Welche Faktoren bestimmen den Stichprobenumfang einer Studie?
  - Ressourcen (Geld + Zeit)
  - Daumenregeln
  - Tradition
  - **Poweranalyse**

Preregistration: Stichprobenplanung

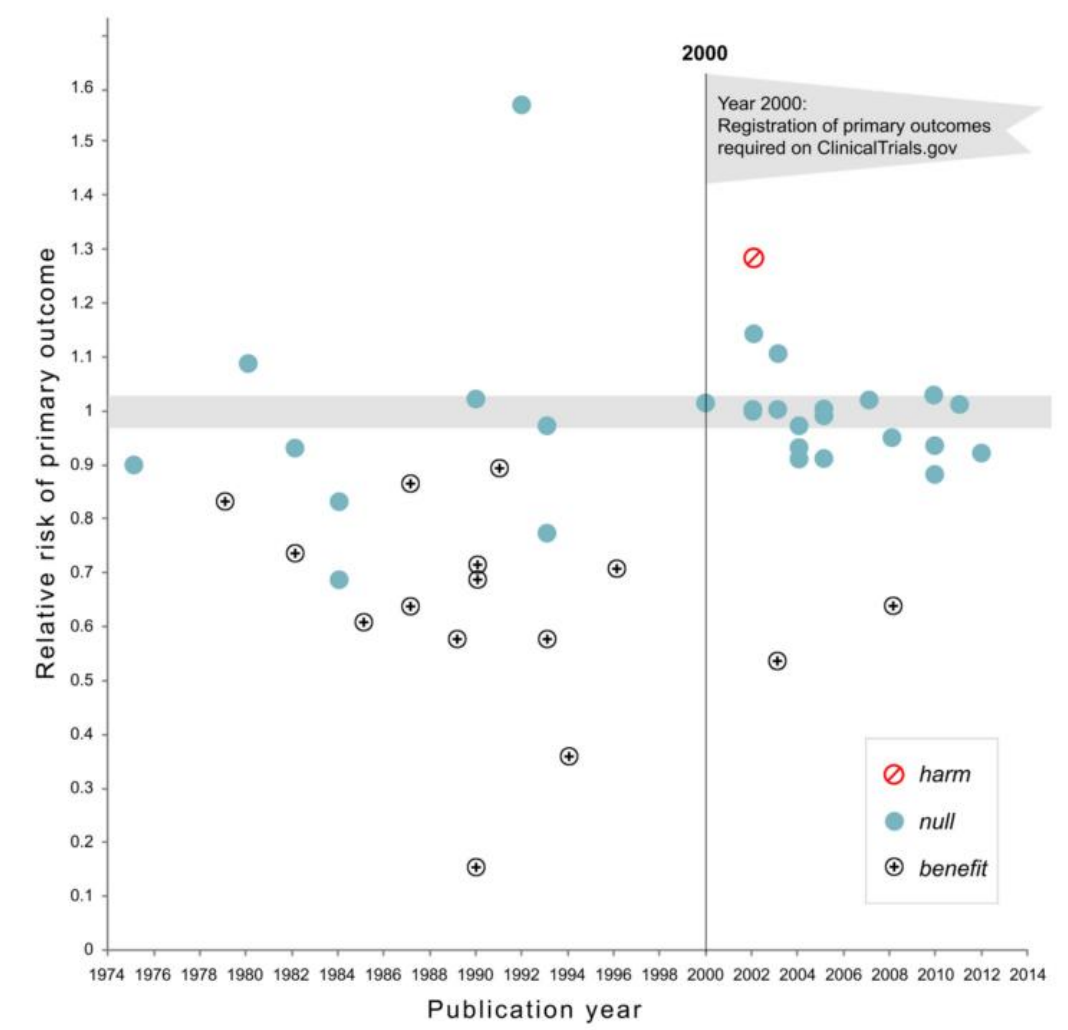
- Poweranalysen
  - Wahrscheinlichkeit den spezifizierten Effekt zu finden, wenn es ihn gibt (**konditional**)
- Warum ist eine Poweranalyse so wichtig?
  - Garantiert Informationsgehalt jeder Studie
  - Mit geringer Power
    - Sind „Null-Ergebnisse“ schwer zu interpretieren
    - Haben statistisch signifikante Befunde eine geringe Replizierbarkeit
  - Mit hoher Power
    - Hohe Wahrscheinlichkeit, Effekte zu beobachten (wenn es sie gibt)
    - „Null-Ergebnisse“ -> Effekt wahrsch. mindestens kleiner als erwartet
  - Ressourceneffizientes Forschen

Ist Preregistration nur sinnvoll für konfirmatorische Forschung?

Nein!

- Zu wissen, dass jemand keine Hypothesen hatte, ist **genau so sinnvoll**
- Preregistration befreit vor jeglichen **inneren und äußeren Zwängen** zur Hypothesentestung

Was passiert wenn plötzlich viele Leute preregisteren?



Preregistration: www.osf.io

- Kollaborative Plattform
- Data Sharing
- Permanentes Repositorium
- Preregistration-Funktion
  - Legt eingefrorene Kopie aller Dateien an, Zeitstempel
  - Wird nach Ablauf festgelegter Zeit öffentlich

Preregistration: www.osf.io