RUHR-UNIVERSITÄT BOCHUM

Die Rolle von Standardisierung in der (meta-analytischen) Interpretation psychologischer Forschung

Malte Elson, Ruhr-Universität Bochum

Seit 2011…

- Phase selbstkritischer Reflexion
  - Das Leben und Sterben von Theorien (Ferguson & Heene, 2012)
  - Exploration und Konfirmation (Wagenmakers et al., 2012)
  - Normen in der Datenerhebung (Lakens & Evers, 2014)
  - Datenauswertung (Wagenmakers et al. 2011)
  - Berichten von Daten (Nuijten et al., 2015)
  - Teilen von Daten (Wicherts, Bakker, & Molenaar, 2011)
  - Publikationspraktiken (Nosek & Bar-Anan, 2012)
  - Peer Review und Qualitätsmanagement (Wicherts et al., 2012)
  - Reproduzierbarkeit und Replizierbarkeit (OSC2015)
  - Wissenschaftskommunikation (Sumner et al., 2014)
  - Belohnungskontingenzen im Wissenschaftssystem (Nosek, Spies, & Motyl, 2012)

Was wissen wir jetzt eigentlich?

- Lehrbücher, Meta-Analysen, Literatur-Reviews, Stellungnahmen basieren alle auf Dekaden kumulativer Forschung

Wie reliabel, wie robust ist dieses Wissen?

Known Knowns & Known Unknowns

- *Known:* Reliabilität psychologischer Forschung ist bedroht durch p-Hacking

> ***p*-Hacking:** eine Klasse von Verhaltensweisen mit dem Ziel, die Wahrscheinlichkeit für statistische Signifikanz in Abwesenheit eines wahren Effekts zu erhöhen

- *Known:* Es gibt effective p-Hacking-Strategien
- *Known:* Psychologen p-hacken ihre Daten (John et al., 2012)
  - Auslassen abhängiger Variablen in Papern (~66%)
  - Exklusion von Daten, nachdem man sie gesehen hat (~43%)
  - Unerwartete Befunde als vorhegesagt berichten (~35%)
- Known: Es gibt Publication Bias

Known Knowns & Known Unknowns

- Techniken zur Biasschätzung i.d.R. ineffektiv zur Identifizierung von p-Hacking
- *Known:* p-Hacking-Strategien
  - Outcome switching
  - Opportune Kontroll-/Moderatorvariablen
  - Multiple Berechnungs-/Auswertungsstrategien
  - *…*
- *Unknown:* Wie sieht ein Korpus an Studien aus, der (teilweise) von diesen Strategien beeinflusst wurde?

Forschungssynthese und Meta-Analysen

- p-Hacking und Publication Bias interagieren
- Hochproblematisch für (quantitative) Forschungssynthese
- TIVA, *p*-uniform, or *p*-curve können (unter bestimmten Bedingungen) Evidenz für Bias in der Literatur aufweisen
- *Unknown*: Zu welchem Ausmaß wurde ist die meta-analytische Effektgröße biased durch Publication Bias, p-Hacking, oder beides?

Forschungssynthese und Meta-Analysen

- **Grundannahme:** Wiederholte Beobachtung eines Zusammenhangs -> Effekt ist Wahrscheinlich echt (exakte Größe verhandelbar)
- Je mehr die **Konsistenz von Ergebnissen** in einem Feld auf **Flexibilität der Methoden** basiert, desto mehr nähert sich die Wahrscheinlichkeit für einen wahren Effekt 0 an.

Standardisierug

- Standardisierte Protokoll = *Manualisierug von Forschung*
  - Wichtige Eigenschaft empirischer Forschung
  - Zeichen für “Reife” einer Disziplin
  - Voraussetzung für psychometrische Objektivität
- Standardisierung bestimmt das **Wie** um die Einflüsse von **Wer, Wo und Wann** auf die Datenqualität (und Inferenzen) zu reduzieren

Standardisierung

- Vorraussetzung für sinnvolle Forschungssynthese
- Kontextualisiert meta-analytische Befunde
  - Objektivität verbreiteter “Paradigmen”
  - Gemeinsames Verständnis von Konstrukten
  - (Fehlen von) Normen, was eigentlich als Evidenz zählt
  - Latente Bedeutung programmatischer Literatur
  - Reproduzierbarkeit, Robustheit und Replizierbarkeit
- (Einige) Informationen in Publikationen verfügbar

Standardisierung

- Homogenität prozeduraler/analytischer Techniken fördert interessantes zu Tage:

  a) **idiosynkratische Flexibilität** und auffällige Abweichungen von Normen in Bereichen, in denen konsistente Normen existieren

  b) **endemische Flexibilität** in Bereichen, in denen es keine Standardisierung gibt

Standardisierung

General Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
XX(X) 1–14

DOI: 10.1177/0956797613480366
pss.sagepub.com
SAGE

Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone

Uri Simonsohn
The Wharton School, University of Pennsylvania

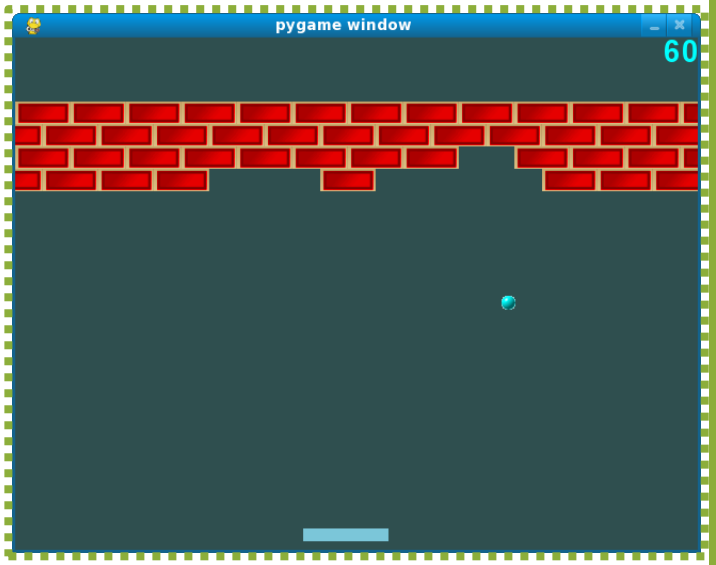
Abstract

I argue that requiring authors to post the raw data supporting their published results has the benefit, among many others, of making fraud much less likely to go undetected. I illustrate this point by describing two cases of suspected fraud I identified exclusively through statistical analysis of reported means and standard deviations. Analyses of the raw data behind these published results provided invaluable confirmation of the initial suspicions, ruling out benign

Beispiele: 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 1



- Klimmt et al. (2007)
- N = 500 Teilnehmer
- 3 Versionen eines Videospiels
  - **Control condition**
  - **Reduzierte Selbstwirksamkeit** (33.3% des Inputs ignoriert)
  - **Reduzierte Kontrolle** (Ball bewegt sich schneller)
- Danach Fragebogen
  - Selbstwirksamkeit (11 items)
  - Kontrolle (9 items)
  - Enjoyment (8 items)

M

-
-
-
-

Rapid Communication

Effectance and Control as Determinants of Video Game Enjoyment

CHRISTOPH KLIMMT, Ph.D.,[1] TILO HARTMANN, Ph.D.,[2] and ANDREAS FREY, Ph.D.[3]

ABSTRACT

This article explores video game enjoyment originated by games' key characteristic, interactivity. An online experiment ($N = 500$) tested experiences of effectance (perceived influence on the game world) and of being in control as mechanisms that link interactivity to enjoyment. A video game was manipulated to either allow normal play, reduce perceived effectance, or reduce perceived control. Enjoyment ratings suggest that effectance is an important factor in video game enjoyment but that the relationship between control of the game situation and enjoyment is more complex.

  - Enjoyment (8 items)

Multiple Analysestrategien: Beispiel 1

- **Effectance:** Causal agency, that is, the perception of receiving immediate, direct feedback on one's action and of influencing the game world
- **Control:** Being in control means to know about the attributes of a situation, to anticipate its dynamics, and to be able to influence it according to one's goals.

Mit meinen Handlungen konnte ich im Spiel bewirken, was ich mir erhofft habe.

Multiple Analysestrategien: Beispiel 1

- **Effectance:** Causal agency, that is, the perception of receiving immediate, direct feedback on one's action and of influencing the game world
- **Control:** Being in control means to know about the attributes of a situation, to anticipate its dynamics, and to be able to influence it according to one's goals.

Es erschien mir, dass das Spiel macht was es will.

Multiple Analysestrategien: Beispiel 1

- **Effectance:** Causal agency, that is, the perception of receiving immediate, direct feedback on one's action and of influencing the game world
- **Control:** Being in control means to know about the attributes of a situation, to anticipate its dynamics, and to be able to influence it according to one's goals.

> Manchmal war mir nicht klar, ob ein Ereignis durch mich ausgelöst wurde, oder ob es auf andere Weise zustande kam.

Multiple Analysestrategien: Beispiel 1

- **Effectance:** Causal agency, that is, the perception of receiving immediate, direct feedback on one's action and of influencing the game world
- **Control:** Being in control means to know about the attributes of a situation, to anticipate its dynamics, and to be able to influence it according to one's goals.

Das Spiel reagierte auf meine Eingaben wie ich es erwartete.

Multiple Analysestrategien: Beispiel 1

- **Effectance:** Causal agency, that is, the perception of receiving immediate, direct feedback on one's action and of influencing the game world
- **Control:** Being in control means to know about the attributes of a situation, to anticipate its dynamics, and to be able to influence it according to one's goals.

> Was immer ich mir vornahm, konnte ich durch meine Handlungen erreichen.

Multiple Analysestrategien: Beispiel 2

- **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 2

- 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)
- …

CRTT: Methodological flexibility

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 2

Ok, und…

Wie sieht es mit der Replizierbarkeit dieser Studien aus?

Wie soll man Meta-Analysen interpretieren, in denen diese Studien enthalten sind?

Kann man Laborforschung zu Aggression überhaupt trauen?

Kann man diese Forschung überhaupt irgendwie sinnvoll zusammenfassen?

Beispiel 3: Inhibitionskontrolle

- Häufigster Test: Go/No-Go Task
  - Jedes Trial besteht aus einem Stimulus: “Go” (respond) oder “No-Go” (do nothing)
  - Bewegung wird aktiviert, muss dann manchmal unterdrückt werden
  - Accuracy = Maß für Inhibitionskontrolle
  - Damit das Paraidgma funktioniert
  - Two important prerequisites for its validity
    - Muss die Aktivität bei jedem Trial ausgelöst werden (No-Go seltener)
    - Aktivität muss unterdrückt werden (Schnelle Trialfolge)

(basierend auf Wessel, 2017)

Example 2: Inhibitory Control

(basierend auf Wessel, 2017)

Fehlende Standardisierung und Forschungsqualität

CRTT	Go/No-Go
- Hohe Variabilität in Prozedur und Analyse - Analysestrategie könnte post hoch gewählt werden - Welche ist die „richtige“ Analysestrategie? - Literatur schwer zu interpretieren Fehlende Standardisierung als Bedrohung der **Objektivität**	- Hohe Variabilität in Prozedure - Designentscheidungen a priori getroffen - Substanzieller Anteil der Studien erfasst nicht das Konstrukt von Interesse - Exklusionskriterium für Meta-Analysen? Fehlende Standardisireung als Bedrohung der **Validität**

Standardisierte Protokolle

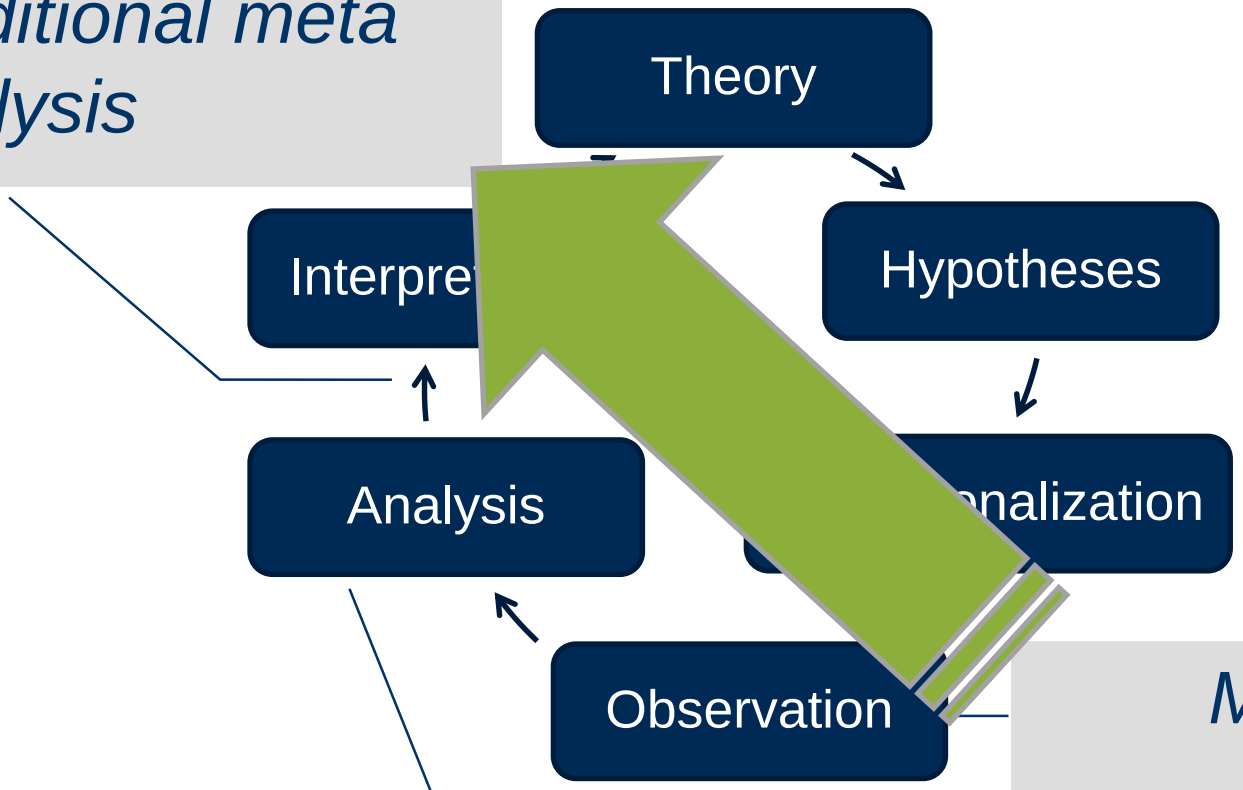
- Entwicklung methodischer „Goldstandards“
  - „Reife“ der Disziplin
  - Leistung der Forschungscommunity
  - Alle relevanten Informationen müssen verfügbar gemacht werden

Standardisierte Protokolle

- Replikationsketten
  - Datenerhebung über verschiedene Labs hinweg
  - Erhöht statistische Power
  - Reduziert Sampling Bias
  - Verbessert Präzision der Schätzung
  - „Erzwingt“ Einigung auf Prozeduren

Meta Method Analysis

Traditional meta analysis



Meta method analysis

HUMAN

TECHNOLOGY

INTERACTION

RUB